Investigating the application of structured representations of unstructured content in personalisation tasks

Aonghus McGovern

A thesis submitted to
University of Dublin, Trinity College
for the degree of
Doctor of Philosophy

March 12, 2019

Declaration

This thesis is entirely the work of the author, except where otherwise stated, and has not been submitted for a degree to any other University. This thesis may be copied and lent to others by the University of Dublin.

Aonghus McGovern

March 12, 2019

Acknowledgements

To James Cogley, who proofed my early drafts and gave me invaluable advice, I will be forever grateful.

To Brendan, Jamie, Lucy, Nicole and Rebekah, the countless discussions we had in the lab will stay with me forever.

To my parents, who raised me and supported me throughout my entire educational journey, I am deeply grateful.

To my supervisor Vincent Wade, who made me into the researcher I am today, goes my sincerest gratitude.

Abstract

For personalisation approaches that analyse unstructured content, a common task is converting that unstructured content to a structured representation. Each structured representation has strengths and weaknesses, and the choice of representation should be made with respect to the personalisation task at hand. However, the way in which the choice of structured representations affects the personalisation that can be performed using that representation has not been clearly articulated. This is because personalisation approaches tend to focus on the success of their chosen personalisation task (e.g. recommendation accuracy) without examining how the characteristics of their chosen structured representation influenced this success. This motivates an investigation of of the characteristics of structured representations in the context of different personalisation tasks. This investigation is the subject of this thesis, and is carried out as a series of experiments. Each of these experiments examines the effect of a single characteristic of structured representations on personalisation performance.

The first experiment investigates how the inability of the Named Entity and Bag-Of-Words representations to capture context limits their ability to fully represent different forms of user expression. This limitation can be overcome by leveraging the contextual information contained in the conceptual hierarchy of an external linguistic resource. The second experiment describes a comparison between the conceptual hierarchies of two different kinds of external linguistic resource: a purely lexical resource (WordNet) and a general knowledge base (DBpedia). The comparison takes the form of an investigation of the ability of each resource to represent the differences between users' descriptions of their interests and knowledge on Twitter with their description of the same characteristics on LinkedIn. The results of this experiment indicate that the DBpedia-based approach is most effective. Another finding of this experiment is that a structured representation's inability to accurately reflect category distinctions affects its ability to provide accurate recommendations. The third experiment investigates this distinction through a series of recommendation tasks spanning multiple domains, user models and recommendation methods. This experiment yields a test to determine whether a structured representation accurately reflects domain category distinctions. Furthermore, this experiment reveals that structured representations that pass this test will facilitate accurate recommendation, while structured representations that do not pass this test will not facilitate accurate recommendation.

The contributions of this thesis consist of indicative guidelines as to the limitations of particular structured representations as well as guidance with respect to the methods for addressing these limitations.

Publications related to this Ph.D.

[1] A. McGovern, A. O'Connor, and V. Wade, "A content analysis of customer support-related tweets," in *2nd European Conference on Social Media ECSM 2015 Porto, Portugal, 9-10 July 2015*, 2015, pp. 332–340.

[2] A. McGovern, A. O'Connor, and V. Wade, "From DBpedia and WordNet hierarchies to LinkedIn and Twitter," in *4th Workshop on Linked Data in Linguistics: Resources and Applications (co-located with ACL-IJCNLP 2015), Beijing, China, 31 July 2015*, 2015.

Contents

1 Introduction 1

1.1 Research Question and Objectives 8

1.2 Research Method 9

1.3 Contributions of this thesis 11

1.4 Overview 11

2 Background 13

2.1 Introduction 13

2.2 Types of structured representation 13

2.3 Types of external resources 15

2.4 Methods for determining weights 16

2.5 Part-Of-Speech tagging and Chunking 19

2.6 A survey of personalisation approaches that apply structured representations of unstructured content 21

2.6.1 Personalisation approaches representing content written by the user . . . 23

2.6.2 Personalisation approaches representing content not written by the user . 26

2.6.3 Personalisation approaches representing content written by the user and content not written by the user 29

2.6.4 The characteristics of structured representations of unstructured content 30

2.6.5 Required properties of the experiments in this thesis 32

2.7 Conclusion 33

3 Examining the Named Entity-based representation 34

3.1 Objective of this experiment 34

3.2 Case description 35

3.3 Research Question of this Experiment 38
3.4 Generating the structured representation 38
3.4.1 Named Entities 39
3.4.2 Bag Of Words 39
3.5 Experiment design 40
3.5.1 Dataset Description 40
3.5.2 Evaluation Method 41
3.5.3 Results 42
3.6 Discussion 53
3.7 Conclusion 54

4 A Comparison of External Linguistic Resources 55

4.1 Objective of this experiment 56
4.2 Case Description 56
4.3 Research Question of this Experiment 57
4.4 Generating the structured representation 58
4.4.1 Keyphrase Identification 59
4.4.2 Linking keyphrases to labels 60
4.4.3 Calculating term weights 66
4.5 Experiment design 67
4.5.1 Dataset Description 68
4.5.2 Evaluation Method 71
4.5.3 Results 77
4.6 Discussion 82
4.7 Conclusion 84

5 Investigating the influence of content domain 86

5.1 Objective of this experiment 87
5.2 Case description 87
5.3 Question of this experiment 88
5.4 Generating the structured representation 89
5.4.1 Creating the different user models for Task 1 89
5.5 Experiment Design 98

5.5.1 Dataset description 98
5.5.2 Generating recommendations 103
5.5.3 Evaluation Method 106
5.5.4 Results 110
5.6 Discussion 116
5.6.1 The relationship between the within-category/between-category test and recommendation accuracy 116
5.6.2 The relationship between similarity scores, diversity and recommendation precision 121
5.6.3 Points to note 121
5.7 Conclusion 122

6 Conclusion **123**
6.1 Summary of thesis contributions 123
6.2 Future work 125
6.3 Concluding Remarks 126

A The WordNet Domains hierarchy with corresponding Dewey Decimal Classification Codes **128**

B Word Sense Disambiguation **138**
B.1 WordNet 138
B.2 DBpedia 140

C The Schulze Method **147**

D Rank-Order Centroids **150**

E Vacancy Categories **155**

F MovieLens Genres **157**

G Music dataset genres **158**

List of Tables

2.1 Example Bag-Of-Words, Named Entity-based and Concept-based structured representations 14

2.2 Two synsets for the word 'bank' with their glosses. 15

2.3 An example RDF predicate. 16

2.4 An example document with term frequency-based weights 18

2.5 A POS-tagged sentence. 20

2.6 A chunked sentence. 20

3.1 Naaman et al.'s forms of user expression 36

3.2 Honeycutt and Herring's forms of user expression 36

3.3 Overlapping categories between Naaman et al. and Honeycutt and Herring. . . . 37

3.4 The combined list of categories from Naaman et al. and Honeycutt and Herring. 37

3.5 Customer support accounts chosen 41

3.6 The top ten entity types, ranked by relative mention. 42

4.1 A chunking of the sample tweet 'What's up @spartan?'. 62

4.2 The WordNet Domains parent domains and DBpedia broader categories for the 'Television', 'Architecture' and 'Literature' subject areas.
77

4.3 The WordNet Domains parent domains and DBpedia broader categories for the 'Television', 'Architecture' and 'Literature' subject areas.
79

4.4 Part 1 and Part 2 Accuracy percentages. 'EWND' denotes the Extended WordNet Domains-based representation and 'DBC' denotes the DBpedia category-based representation. 81

4.5 Part 1 and Part 2 Accuracy percentages. ‘EWND’ denotes the Extended WordNet Domains-based representation and ‘DBC’ denotes the DBpedia category-based representation. 81

4.6 Error rate percentages. ‘EWND’ denotes the Extended WordNet Domains-based representation and ‘DBC’ denotes the DBpedia category-based representation. . 81

5.1 Comparison of the Schulze Method with other voting methods [111]. 94

5.2 Movie and Music dataset statistics for AlchemyAPI. 102

5.3 Movie and Music dataset statistics for Babelfy. 102

5.4 News dataset category similarity scores. 110

5.5 Vacancy categories with higher ‘Between category’ similarity than ‘Within category’ similarity. 111

5.6 Movie genres with higher ‘Between category’ similarity than ‘Within category’ similarity, using AlchemyAPI. 111

5.7 Movie genres with higher ‘Between category’ similarity than ‘Within category’ similarity, using Babelfy. 112

5.8 Music genres with higher ‘Between category’ similarity than ‘Within category’ similarity, using AlchemyAPI. 112

5.9 Music genres with higher ‘Between category’ similarity than ‘Within category’ similarity, using Babelfy. 113

5.10 News recommendation precision percentages. 113

5.11 Job recommendation precision percentages. 113

5.12 Movie and Music Recommendation precision, using AlchemyAPI. 114

5.13 Movie and Music Recommendation precision, using Babelfy. 114

5.14 Logistic Regression results for the relationship between similarity scores, diversity and recommendation accuracy for the news reoommendation task. 114

5.15 Logistic Regression results for the relationship between similarity scores, diversity and recommendation accuracy for the job recommendation task. 114

5.16 Logistic Regression results for the relationship between similarity scores and recommendation accuracy for the music recommendation task with AlchemyAPI. 115

5.17 Logistic Regression results for the relationship between similarity scores and recommendation accuracy for the music recommendation task with Babelfy. . . 115

5.18 Logistic Regression results for the relationship between similarity scores and recommendation accuracy for the movie recommendation task with AlchemyAPI. 116

5.19 Logistic Regression results for the relationship between similarity scores and recommendation accuracy for the movie recommendation task with BabelNet. . 116

5.20 Non-unique subjects in the Vacancy domain. 118

5.21 Non-unique subjects in the Movie domain for the AlchemyAPI analysis tool. . . 119

5.22 Non-unique subjects in the Movie domain for the Babelfy analysis tool. 119

5.23 Non-unique subjects in the Music domain for the AlchemyAPI analysis tool. . . 120

5.24 Non-unique subjects in the Music domain for the Babelfy analysis tool. 120

B.1 Table 1 from Magnini et al.'s WSD paper. 139

B.2 Information required for x^2 independence testing. 141

B.3 Sample values for the cells in Table B.2 142

B.4 Mihalcea and Csomai's results from comparing the tf-idf, x^2 and keyphraseness ranking methods. 143

B.5 Possible definitions for the word 'cone'. 144

B.6 Mihalcea and Csomai's results from comparing knowledge- and feature-based WSD algorithms. 146

C.1 Example ballots. 149

C.2 A pairwise comparison matrix for Table C1. 149

C.3 A pairwise comparison matrix of strongest paths for Table C.2. 149

D.1 The Rank-Sum, Rank-Reciprocal and Rank-Order Centroid weights for four ranked items. 151

D.2 Bannon and Barrett's results from comparing the Rank-Reciprocal and Rank-Order Centroid weights. 154

E.1 Job vacancy categories 156

F.1 MovieLens genres 157

G.1 Music dataset genres 158

List of Figures

1.1 An example term-list representation. 2

2.1 A screenshot of the Wikipedia category page for 'Semantic Web'. 17

2.2 Two example documents with term frequencies. 19

2.3 An example tweet from Mendes et al. 28

3.1 A tweet containing a link un-related to customer support. 40

3.2 A spam tweet. 40

3.3 A tweet to the AT&T account. 41

3.4 A tweet to the Microsoft account. 41

3.5 An information sharing tweet to the Microsoft account. 43

3.6 A tweet to the TMobile UK account. 43

3.7 An information sharing tweet to the Eircom account. 43

3.8 An opinion to the Orange UK account. 44

3.9 An opinion to the AT&T account. 44

3.10 An opinion to the Comcast account. 44

3.11 A statement to the Lexmark account. 45

3.12 A statement to the Verizon account. 45

3.13 A statement to the HP account. 45

3.14 A 'Me Now' tweet to the Microsoft account. 46

3.15 A 'Me Now' tweet to the Verizon account. 46

3.16 A 'Me Now' tweet to the Eircom account. 46

3.17 A 'Question to followers' tweet to the AT&T account. 47

3.18 A 'Question to followers' tweet to the Comcast account. 47

3.19 A 'Question to followers' tweet to the Verizon account. 47

3.20 A 'Presence Maintenance' tweet from the Sony account. 48

3.21 A ‘Presence Maintenance’ tweet from the Dell account. 48
3.22 A ‘Presence Maintenance’ tweet from the HP account. 48
3.23 An ‘About Addressee’ tweet to the Comcast account. 48
3.24 An ‘About Addressee’ tweet to the Dell account. 48
3.25 An ‘About Addressee’ tweet to the Microsoft account. 49
3.26 An ‘Anecdote (me)’ tweet to the Dell account. 49
3.27 An ‘Anecdote (me)’ tweet to the Microsoft account. 49
3.28 An ‘Anecdote (me)’ tweet to the Yahoo account. 50
3.29 An ‘Anecdote (others)’ tweet to the Orange UK account. 50
3.30 An ‘Anecdote (others)’ tweet to the LG account. 50
3.31 An ‘Anecdote (others)’ tweet to the Yahoo account. 50
3.32 A ‘Self promotion’ tweet from the O2 account. 51
3.33 A ‘Self promotion’ tweet from the Comcast account. 51
3.34 A ‘Self promotion’ tweet from the Verizon account. 51
3.35 An ‘Exhort’ tweet from the O2 account. 52
3.36 An ‘Exhort’ tweet to the AT&T account. 52
3.37 An ‘Exhort’ tweet to the Vodafone account. 52
3.38 A ‘Media Use’ tweet to the Nokia account. 53
3.39 A ‘Media Use’ tweet to the Microsoft account. 53
3.40 A ‘Media Use’ tweet to the Comcast account. 53

4.1 A paragraph from the author’s LinkedIn summary. 69
4.2 A truncated version of the author’s LinkedIn skill list. 69
4.3 A truncated version of the author’s LinkedIn interest list. 69
4.4 A truncated version of the author’s LinkedIn course list. 69
4.5 A textual description of the author’s M.Phil degree. 70
4.6 A paragraph from the LinkedIn description of an internship undertaken by the author. 70
4.7 A paragraph from the description for the ‘Computational Linguistics’ LinkedIn group. 70
4.8 Example term prominence comparison and example recommendation 74
4.9 Two different rankings of a set of 5 terms. 84

5.1 Two different term-frequency lists for a collection of 4 terms. 93

5.2 The Additive-Aggregated User Model generated from the term-frequency lists in Figure 5.2. 93

5.3 The Rank-Aggregated User Model generated from the term-frequency lists in Figure 5.2. ‘Term 3’ and ‘Term2’ are tied. 93

5.4 A graph of number of authors per number of publications on DBLP. 100

5.5 Sample weighted term-lists and the associated vectors. 104

D.1 An example of a normal (blue) and uniform (green) distribution over the range [0,1]. 153

Chapter 1

Introduction

Personalised services are those that "build a model of the individual user and apply it for adaptation[1] to that user" [4]. Personalisation offers benefits in various applications, including customer support [5], web search [6] and recommender systems for items such as news [7]. In [5], Steichen et al. leverage information in user-generated content such as fora, blogs etc. to supplement professionally authored content in order to provide personalised responses to customer support queries. Hecht et al. leverage users' social network on Facebook to provide personalised search results [6]. For example, if a user posts the question `'On Verizon, will I have to pay roaming charges if I use my cell phone in Hawaii?'`, Hecht et al.'s system will inform the user that some of their Facebook friends currently or formerly live in or near Hawaii. Abel et al. analyse the content of users' tweets, as well as news articles linked in those tweets, to recommend news articles. A user's interests are obtained from tweets that user has written and articles they have read. These interests are then used to identify news articles of interest to that user. Although the application areas of such approaches vary, they share a common fundamental task: the use of Natural Language Processing (NLP) methods to generate structured representations of unstructured content. Writing for Forbes, Robert Malone describes unstructured content as "data that is free-form that can be from written notes, word documents, e-mail, reports, news feeds, etc." [8].

[1] In [3], Brusilovsky defines adaptation as the quality of a system behaving differently for different users.

Structured representations of unstructured content[2] capture relevant information from that unstructured content [9], [10]. In the context of personalisation, information is ‘relevant’ if it facilitates the particular kind of personalisation being provided. A commonly employed structured representation format is the term-list. These terms can be single- or multi-word. Optionally, terms can be assigned a numerical value. These values represent the weight assigned to each term, indicating that term’s prominence in the list. This thesis employs the term-list form of structured representation. An example of such a list can be seen in Figure 1.1.

```
[‘Sports’:0.5, ‘Politics’:0.25, ‘Science’:0.25]
```

Figure 1.1: An example term-list representation.

Personalisation approaches employ structured representations of unstructured content in one of three ways:

- To represent content written by the user to whom the personalisation approach is adapting - Hecht et al.’s approach aims to find users who can answer questions the user has written. These authors use a structured representation consisting of Named Entities (i.e. people, places, organisations etc.) to represent users’ unstructured questions in order to identify information that allows their application to find users who can answer these questions. Consider the example sentence on the previous page. The Named Entity-based representation identifies the location Hawaii from this sentence. Hecht et al.’s application can then find relevant users for this question by searching the user’s Facebook friends list for users with Hawaii in their location field.

- To represent content created by someone other than the user to whom the personalisation approach is adapting - Steichen et al.’s approach aims to use unstructured content from sources such as fora and blogs in responding to user customer support issues. These authors use a structured representation consisting of concepts to represent this unstructured content in order to find information relevant to the user’s query. Steichen et al.’s structured representation is created by mapping unstructured content to items in external knowledge bases such as DBpedia[3]. These authors’ approach creates a personalised composition of these items, based on the user’s query.

[2] By ‘unstructured content’, the author refers to text only and not to data such as sensor information, speech etc.

[3] DBpedia is a machine-readable representation of the information contained in Wikipedia info boxes[11]. For a full description of DBpedia please see Section 2.3.

- A combination of the two previous points - Abel et al.'s approach aims to use users' tweets - and the links those tweets contain - to recommend news articles. This creates a more complete representation of users' news interests than simply using just tweets or just linked news articles. These authors compare three structured representations: (i) One consising of Named Entities (ii) One consisting of concepts (iii) One consisting of hashtags. News articles to be recommended are represented as lists of: (i)Named Entities (ii) Concepts (iii) Words (i.e. the text of the article). Abel et al. recommend articles for users based on the similarity between the list representing the user model and the lists representing the news articles. These authors find that the Named Entity-based representation provides the most accurate news recommendations.

Personalisation approaches that generate structured representations of unstructured content are common. Approximately 31% of the papers accepted to the 2016 Conference on User Modeling, Adaptation and Personalization (UMAP) performed personalisation in this manner [12]. For example, Huang et al. generate structured representations of learning material in order to determine how well students are learning that material [13]. Herzig et al. generate structured representations of social media customer support discussions in order to predict customer satisfaction [14]. Oyibo et al. generate structured descriptions of survey comments in order to identify gender differences in perception of User Interfaces [15][4].

The choice of structured representation naturally affects the personalisation that can be performed using that representation. Consider Hecht et al.'s approach [6]. Suppose these authors had employed a structured representation that did not contain locations. This would mean their application would have been unable to identify the location 'Hawaii' in the question `'On Verizon, will I have to pay roaming charges if I use my cell phone in Hawaii?'`. Hecht's approach would thus have been hindered in finding relevant users for this question. By the same token however, the fact that Hecht et al.'s chosen structured representation focuses solely on Named Entities means there is certain information that is not captured. The Named Entity representation is content-based, meaning it is derived solely from the text being analysed. In Hecht et al.'s approach, if no Named Entities are identified in a question, that question is not answered. Importantly, this means that Hecht et al.'s approach can only provide answers to questions where the answer can be derived from the text of the question itself. If the answer to the question depends on a broader context that is referenced indirectly in the question, this approach will flounder. Hecht et al. report that their approach provided answers for 26.7% of the questions that were identified. In other words,

[4] The five sessions in UMAP 2016 contained a total of 26 papers. With the three papers mentioned in this paragraph and five others ([16], [17], [18], [19], [20]) the ratio is 8/26, or approximately 31%.

nearly three quarters of posted questions went unanswered. This demonstrates that, in most cases, the Named Entity-based representation was unable to answer a user question. However, exactly how the Named Entity-based representation's inability to capture context relates to its inability to answer user questions is not examined.

There is currently an absence of guidelines as to the how the choice of structured representation affects the personalisation that can be provided using that representation. On the one hand, approaches such as Abel et al. [7] and Hecht et al. [6] employ the Named Entity-based representation to provide personalisation. On the other hand, researchers such as Michelson and Macskassy and Kapanipathi et al. argue that simply identifying Named Entities is not enough to get a complete view of the user [21] [22]. These authors argue that identified Named Entities must be grouped together using contextual information from the conceptual hierarchy in an external linguistic resource. This contextual information facilitates the creation of a categorised description of the user's information. For example, if a user is interested in 'Apple, Inc.' and 'Microsoft Inc.' it can be inferred that they may be interested in technology companies. However, approaches such as Hecht et al. and Abel et al. dn not examine how the limitations of the Named Entity based-representation affect personalisation performance. Approaches such as Michelson and Macskassy and Kapanipathi et al. do not clearly articulate the limitations of the Named Entity-based representation and whether the conceptual hierarchy in their chosen external linguistic resource addresses these limitations.

This raises the question of how to determine which external linguistic resource to employ. Linguistic resources can be broadly divided into purely lexical resources (e.g. WordNet) and general knowledge bases (e.g. DBpedia) [23]. Different linguistic resources will yield different categorisations. Consider the word 'law', as defined by WordNet and DBpedia. Each resource has a similar definition for this word. In WordNet, 'law' is defined as 'the collection of rules imposed by authority'[5]. In DBpedia, 'law' is defined as 'a system of rules that are enforced through social institutions to govern behavior'[6]. However, each of these resources defines the more general form of 'law' quite differently. In WordNet, the more general form of 'law' is 'collection', defined as 'several things grouped together or considered as a whole'. In DBpedia, the more general form of 'law' is 'justice', which is defined as 'the legal or philosophical theory by which fairness is administered'. Bentivogli et al. argue that the validity of a linguistic resource can be assessed by examining that resource's *social relevance* [24]. The social

[5] The WordNet definition for 'law' can be found here: `http://wordnetweb.princeton.edu/perl/webwn?s=law&sub=Search`

[6] The Wikipedia page for 'law' can be found here: `https://en.wikipedia.org/wiki/Law`

relevance of a resource describes the extent to which that resource reflects the way in which information is represented in society in general. For example, a resource in which 90% of the categories relate to History would have low social relevance, as it represents predominantly one subject. When determining whether the context provided by an external linguistic resource is appropriate for a personalisation task, the social relevance of that resource must be considered.

Content domain must also be considered when leveraging an external linguistic resource. Different personalisation approaches have defined 'content domain' differently. Some of these ways are:

- The service in which the content resides. For example, Alanazi et al. describe cross-domain recommendation, where users' tweets are used to bootstrap recommendations in a news recommender system [25].
- The subject area of the content. For example, Paris and Wan describe an analysis of social media language use between a community in the science domain and the social services domain [26].
- As a distinct category of item. For example, Arguello et al. define news articles, job postings etc. as being separate domains [27].

In this thesis 'content domain' is defined in the same way as Arguello et al., for example news, jobs, etc. A structured representation can yield substantially different results in the same personalisation task, depending on the domain of the content being analysed in that task. Consider for example the task of recommending items for users. While a structured representation may facilitate effective recommendation for one content domain, the same structured representation may hinder recommendation in another content domain. This occurs when a structured representation does not accurately reflect domain category distinctions. In order to determine whether a structured representation accurately reflects domain category distinctions, it is necessary to compare between-category and within category similarity. If a structured representation yields higher between-category similarity than within-category similarity, this representation will not facilitate effective recommendation. For example, consider the task of performing recommendation in the movie domain. Suppose that the chosen structured representation yields higher average similarity for items in the Action category to items in the Comedy category than it does for items in the Action category to other items in the Action category. This representation will not facilitate effective recommendation, since it represents spurious similarities between items. In this scenario a person who has liked numerous Action films could be incorrectly recommended Comedy films, and vice versa.

The author characterises the choice of structured representations in personalisation approaches such as [5], [6] and [7] according to the following dimensions:

1. The type of information the structured representation contains e.g. Named Entities such as people and places, concepts such as politics and sport etc.[7]

2. The external linguistic resource employed in generating the structured representation e.g. a purely lexical resource such as WordNet or a general knowledge base such as DBpedia.

3. The domain of the content being analysed in the personalisation task.

Note that these characteristics are not considered by focusing solely on the success of a particular personalisation task. Yet, each of these characteristics affects personalisation performance. Hecht et al. report that nearly three quarters of user questions went unanswered. However, these authors do not examine how the limitations of their chosen structured representation (the Named Entity-based representation) contributed to this result. Approaches such as Abel et al. [28] and Orlandi et al. [29] employ conceptual hierarchies in external linguistic resources to address these limitations. Abel et al. evaluate their approach in a tag prediction task. Orlandi et al. generate a list of interests for participants and ask participants to subjectively rate the accuracy of the list. Recall from the previous discussion the differences in categorisation between different kinds of external linguistic resource. Neither Abel et al. nor Orlandi et al. compare their approach against an approach that uses a different kind of external linguistic resource. This means the efficacy of each approach's chosen resource cannot be reliably determined. Guo et al. compare multiple different structured representations in a job recommendation task, reporting recommendation precision for each representation [30]. These authors find that different representations yield different precision results. This indicates that certain representations represent the job items more accurately than others. However, these authors do not fully examine the reasons *why* one structured representation yields better results than another.

This thesis examines structured representations of unstructured content that differ with respect to the aforementioned three characteristics. The examination will focus on the following unexplored areas:

1. An investigation of the limitations of the Bag Of Words and Named Entity-based representations. On the one hand, Abel et al. find that the Named Entity-based representation

[7] In this thesis, the word 'Concept' is used to refer to the abstraction of an object or idea whereas the term 'Named Entity' is used to refer to a single entity. For example, 'Person' is a concept while 'Barack Obama' is a Named Entity.

is the most effective representation for a specific task (recommending news articles). On the other hand, Michelson and Macskassy and Kapanipathi et al. argue that, to get a more complete picture of the user, contextual information must be provided by leveraging the conceptual hierarchy in an external linguistic resource must be employed. This raises an issue with regard to what the limitations of content-based representations such as Named Entities are, and whether approaches such as Michelson and Macskassy and Kapanipathi et al. address these limitations. This thesis investigates this issue by examining how the inability of the Named Entity and Bag Of Words representations to capture contextual information limits their ability to fully represent different forms of user expression.

2. A comparison of different types of external linguistic resource in representing user interests and knowledge. Recall the preceding discussion on the different categorisation hierarchies provided by different types of external linguistic resource. This raises the issue of choosing an external linguistic resource, to provide the context that is abset in content-based representations such as Named Entities and Bag Of Words. Recall also the preceding discussion on how the criterion of social relevance can aid this choice. This thesis investigates this issue by examining how the social relevance of DBpedia and WordNet affects their ability to represent user Interests and Knowledge.

3. An investigation of the influence of content domain on the efficacy of structured representations in personalisation tasks. Guo et al. compare different structured representations in a job recommendation task but do not fully investigate the reasons for differences in performance between structured representations. This raises the issue of why one structured representation yields better recommendations in a particular content domain than another. Recall the preceding discussion on assesssing the suitability of a structured representation by comparing between-category and within-category similarity. This thesis investigates this issue through a recommendation experiment spanning multiple content domains, user models and recommendation methods.

There are naturally other characteristics of structured representations of unstructured content. These are the analysis method employed to generate the structured representation and the way in which items in the structured representation are weighted. Each of these considerations is discussed further in the following subsections.

The analysis method employed to generate the structured representation

Both Michelson and Macskassy and Kapanipathi et al. employ the Named Entity-based representation to represent user interests. However, Kapanipathi et al. use a pre-built tool, Zemanta, whereas Michelson and Macskassy implement Named Entity identification as part of their approach.

Weighting
Different personalisation approaches employ different weighting schemes according to their requirements. For example, Hauff and Houben weight according to term frequency [31], whereas Orlandi et al. combine frequency and regularity of occurrence [29].

The above characteristics are approach-specific. As such, format and weighting considerations are not the focus of this thesis, but are addressed in its individual chapters as required.

1.1 Research Question and Objectives

This section describes the research question of this thesis, as well as the objectives for addressing this question.

The research question is as follows:

How does the choice of structured representation[8] *affect the personalisation that can be provided using that representation?*

The task of addressing this question is divided into the following objectives:

1. An investigation of how the inability of representations such as Named Entities and Bag-Of-Words to capture context limits their ability to represent different forms of user expression.
2. A comparison of the efficacy of two different kinds of external linguistic resource in providing the context that is absent in representations such as Named Entities and Bag-Of-Words. This comparison will examine the social relevance of each resource; and how this social relevance affects each resource's ability to represent user interests and knowledge.
3. An investigation of the influence of content domain on the efficacy of structured representations. This investigation examines the between-category/within-category distinction discussed in the previous section. The examination takes the form of an experiment spanning multiple content domains,user models and recommendation methods.

[8] Where the choice of structured representation is characterised according to the dimensions given on page 6.

Objective 1 assesses the specific limitations of representations such as Bag Of Words and Named Entities. Objectives 2 and 3 examine different considerations that must be taken into account when attempting to overcome these limitations by employing a conceptual hierarchy in an external linguistic resource, namely social relevance and the between-category/within-category distinction.

1.2 Research Method

The first step in addressing this research question is performing a review of various different structured representations and the NLP methods used to generate those representations. This review concludes with a survey of approaches employing structured representations of unstructured content for the purposes of personalisation. This survey highlights unanswered questions with regard to how the choice of structured representation affects personalisation performance. This review is followed by an experiment investigating how the inability of representations such as Named Entities and Bag-Of-Words to capture context limits their ability to represent different forms of user expression. A common approach for addressing this limitation is leveraging the contextual information contained in the conceptual hierarchy of an external linguistic resource. The second experiment in this thesis compares two different kinds of external linguistic resource. Finally, an experiment is performed to investigate the way in which the domain of the content being analysed in a personalisation task affects the efficacy of structured representations in that task.

For the purposes of this thesis, the efficacy of a structured representation for a personalisation approach is measured in terms of the extent to which the representation facilitates the approach in performing personalisation. For example, in Abel et al.'s study, the Named Entities were more effective for the task of news recommendation than concepts or hashtags [7]. This is because Named Entities yielded recommendations of higher accuracy than concepts or hashtags representations. Michelson and Macskassy and Kapanipathi et al. argue that the Named Entity representation is ineffective in capturing user interests. This is because, these authors argue, Named Entities do not capture categories of user interest [21] [22].

Recall however that the objectives described in the previous section are not concerned with specific personalisation approaches. Instead, they are concerned with the process of personalisation. For example, it is not the goal of this thesis to determine the best way to recommend news articles. Recall also that the discussion in the previous section focused on the lack of clarity about how the choice of structured representation affects the personalisation that structured representation could provide. The key aspect of this discussion was that the nature of the relationship between the choice of structured representation and personalisation performance had not been examined. For example, Abel et al. find that the Named Entity-based representation yields more accurate news recommendations than the concept-based or hashtag-based representations, but do not fully examine *why* this occurs. The research method employed in this thesis must therefore seek to highlight this relationship.

Based on the preceding paragraphs, the research method of this thesis must provide a means of investigating the nature of the relationship between the choice of structured representation and personalisation performance. This investigation must not be specific to a particular personalisation approach, so that its findings are generalisable. The author thus determines that the most appropriate research method for this thesis is a series of experiments on individual representative cases of specific phenomena. These experiments will perform in-depth examinations of each of these cases in order to fully elucidate mechanisms of effect in a way that will be generalisable for the phenomena represented by these cases.

It should be noted that the number of participants for some of the experiments in this thesis are quite small, making the findings indicative rather than definitive. However, as Flyvbjerg states:

``from both an understanding-oriented and an action-oriented perspective, it is often more important to clarify the deeper causes behind a given problem and its consequences than to describe the symptoms of the problem and how frequently they occur. Random samples emphasizing representativeness will seldom be able to produce this kind of insight; it is more appropriate to select some few cases chosen for their validity.''[32][9]

The author argues that the validity of the experiments described in this thesis derives from the unique qualities of the participants. These qualities allow for a full investigation of the phenomena being examined in the experiments of this thesis.

[9] Flyvbjerg makes this point with respect to choosing cases for case studies. However, Flyvbjerg defines a case study as an experiment on a small number of 'valid' cases. Therefore, although the case study method is not being applied in this thesis, Flyvbjerg's point on the validity of participants is still relevant here.

1.3 Contributions of this thesis

The contributions in this section are presented in an order different to that in which they appear in Chapters 3-5. This is done so as to list the contributions in the order of their significance. The contributions of this thesis are as follows:

1. An identification and demonstration that if structured representations of unstructured content do not reflect the category distinctions of that content's domain, recommendation performance will be substantially degraded. This property has been demonstrated across different content domains, user models and recommendation methods. This thesis also provides a straightforward method for testing whether a structured representation reflects domain category distinctions.

2. An indication that the greater social relevance of the DBpedia category hierarchy allows it to provide more accurate representations of user interests and knowledge than the WordNet Domains hierarchy.

3. An indication as to how the inability of representations such as Named Entities and Bag-Of-Words to capture contextual information limits their ability to fully describe different kinds of user expression.

These contributions highlight key limitations in content-based representations, as well as providing important considerations that must be taken into account when leveraging an external linguistic resource to overcome these limitations.

1.4 Overview

Chapter 2 provides a review of the application of NLP methods to generate structured representations of unstructured content for the purposes of personalisation. This review discusses the following:

1. Different types of structured representations generated from unstructured content

2. External linguistic resources that are employed in generating these structured representations

3. Part-Of-Speech tagging - the process of labeling words in text as nouns, verbs etc. - and chunking - which uses Part-Of-Speech tags to group words in a text according to the roles they play e.g. noun phrase, verb phrase etc.

4. A survey of research approaches employing structured representations of unstructured content for the purposes of personalisation. This discussion highlights gaps in the current state of the art that will be addressed by this thesis.

Chapter 3 describes an experiment examining how the inability of the Named entity and Bag-Of-Words representations to capture context limits their ability to fully represent different forms of user expression. A paper related to this work is published in [1]. A common method for overcoming this limitation is leveraging the contextual information contained in the conceptual hierarchy in an external linguistic resource. Such resources can be broadly classified into two types, purely lexical resources and general knowledge bases [23]. **Chapter 4** describes an experiment comparing a representative resource of each type. WordNet is chosen as an example of a purely lexical resource while DBpedia is chosen as an example of a general knowledge base. The comparison takes the form of an investigation of the difference between the way users describe their interests and knowledge through their Twitter posts and the way they describe the same characteristics through their profile information on LinkedIn. A paper related to this work is published in [2]. This experiment finds that the DBpedia-based structured representation appears to be the most effective in this task. This experiment also finds that if a structured representation does not accurately represent category distinctions, recommendation accuracy can suffer. **Chapter 5** describes an experiment investigating the ability of a structured representation to accurately reflect domain category distinctions, and how this ability relates to recommendation accuracy. This experiment takes the form of a series of recommendation tasks spanning multiple domains, user models and recommendation methods. **Chapter 6** summarises the contributions of this dissertation, discusses areas for future work and concludes.

Chapter 2

Background

2.1 Introduction

As discussed in Chapter 1, various personalisation approaches apply NLP methods to unstructured content in order to generate structured representations of that content. This chapter discusses the different considerations involved in generating these structured representations. Recall also that this thesis focuses on the term-list format of structured representation.**Section 2.2** discusses different types of term-list structured representation (referred to as simply 'structured representation' from now on). **Section 2.3** discusses the different types of external linguistic resource that have been employed in order to generate structured representations. **Section 2.4** discusses different methods for determining the weights of terms in structured representations. **Section 2.5** describes different NLP techniques that will be applied in the case studies described in this dissertation. **Section 2.6** surveys various different personalisation approaches that employ structured representations. This survey motivates the examination of different structured representations conducted in this thesis. **Section 2.7** concludes.

2.2 Types of structured representation

As described in the previous section, this chapter discusses term-list structured representations of unstructured content. Different personalisation applications define the word 'term' differently. One of the most basic representations directly uses the words that appear in the unstructured content. This is known as the *Bag-Of-Words* representation. Often, 'stop' words are not considered. These are words such as 'and', 'the' etc. that do not capture information that is relevant for personalisation, but instead perform a specific lexical function within a text. The Bag-Of-Words model has been applied for the purposes of automatically inferring a user's location based on texts they have written [33] and approximating a user's bookmarking profile [31].

Certain approaches aim to enrich words in text by linking these words with terms in an external linguistic resource. The types of terms to which words are linked are broadly classified in the following subsections.

Named Entities

Named Entities are terms such as people, places and organisations. Examples of personalisation approaches employing the Named Entity-based representation include searching social web streams for information relating to real-time incidents [34], filtering documents for users based on the documents' geographical scope (i.e. whether the document content is relevant to the user's location) [35] and providing personalised product recommendations to shoppers [36].

Concepts

In this thesis, the word 'Concept' is defined as referring to the abstraction of an idea e.g. 'car', 'balloon', 'politics', 'sports', etc. Examples of personalisation approaches employing the concept-based representation include facilitating users in finding content and people related to their area of expertise [37] and grouping users into distinct communities based on their opinions with regard to certain topics e.g. movies [38].

An example of each of the Bag-Of-Words, Named Entity-based and concept-based representations can be seen in Table 2.1, using the following sentences: 1. `I love Barack Obama.` 2. `I hate Rush Limbaugh.` 3. `I am in the Democratic Party.`

Bag-Of-Words
Love
Barack
Obama
Hate
Rush
Limbaugh
Democratic
Party

Named Entity-based
Barack Obama
Rush Limbaugh
Democratic Party

Concept-based
American Politics
Politics
Liberalism

Table 2.1: Example Bag-Of-Words, Named Entity-based and Concept-based structured representations

2.3 Types of external resources

Chiarcos et al. divide external linguistic resources into two categories: (i) Strictly lexical resources (ii) General knowledge bases [23]. An example of each type of resource is given in the following subsections.

WordNet (Purely lexical resource)
WordNet is "an online lexical database designed for use under program control. English nouns, verbs, adjectives, and adverbs are organized into sets of synonyms, each representing a lexicalized concept. Semantic relations link the synonym sets" [39]. WordNet groups English words into lexical units called synsets. As a word can have multiple meanings, a word can have multiple synsets. Each synset represents a single meaning of the word. Each synset has an associated 'gloss' i.e. a description of the synset's meaning. Two example synsets for the word 'bank' can be seen in Table 2.2. Synsets are linked by a variety of lexical relations e.g. synonymy, antonymy etc. WordNet has been applied in personalisation studies such as Semeraro et al., for recommending papers in academic conferences [40] and Bellekens et al. in their Sensee framework that provides personalised access to television content [41].

Sense	**Gloss**
1	sloping land (especially the slope beside a body of water)) "they pulled the canoe up on the bank";"he sat on the bank of the river and watched the currents"
2	(a financial institution that accepts deposits and channels the money into lending activities) "he cashed a check at the bank";"that bank holds the mortgage on my home"

Table 2.2: Two synsets for the word 'bank' with their glosses.

DBpedia (General knowledge base)
DBpedia is "a community effort to extract structured information from Wikipedia and to make this information available on the Web." [11]. The DBpedia knowledge base represents Wikipedia information using the Resource Description Framework (RDF) model. With respect to RDF, a 'resource' is a web resource that is referred to by means of a Uniform Resource

Identifier (URI)[1]. RDF encodes relations between resources in the form of subject-predicate-object triples which state that "some relationship, indicated by the predicate, holds between the things denoted by the subject and object of the triple" [42]. An example RDF triple can be seen in Table 2.3. This triple states that the 'Semantic Web' DBpedia category is related by the Simple Knowledge Organization System (SKOS)[2] 'broader' relation to the DBpedia 'Data management' category. The 'broader' relation links its subject to an object that is semantically more general [43].

Subject	**Predicate**	**Object**
`http://dbpedia.org/resource/Category:Semantic_Web`	`http://www.w3.org/2004/02/skos/core#broader`	`http://dbpedia.org/resource/Category:Data_management`

Table 2.3: An example RDF predicate.

DBpedia categories are derived from Wikipedia category pages, which group Wikipedia pages on similar subjects. A Wikipedia category page lists:

- The category's subcategories.
- An alphabetised list of pages in the category.
- The category's super-categories. These appear in a space-separated list at the bottom of the page after the text 'Categories:'.

Figure 2.1 shows the Wikipedia category page for 'Semantic Web'[3]. Note that 'Data management' can be seen in the super-categories list.

DBpedia has been applied by Yang et al. to provide personalised web search [44] and by Tao et al. in their TUMS framework that models users based on their tweets [45].

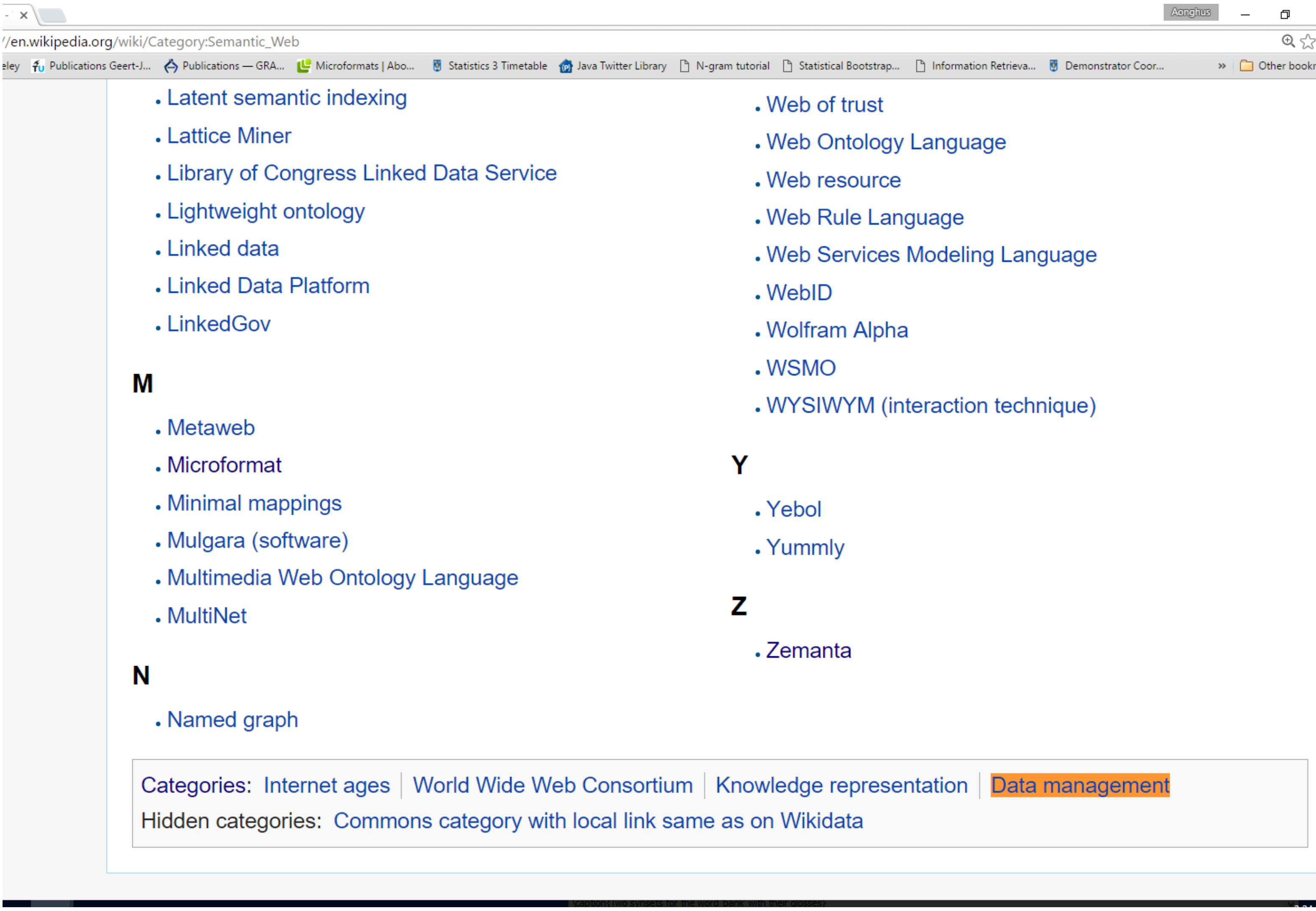
2.4 Methods for determining weights

This section describes two of the most common methods for determining the weights of terms in structured representations of unstructured content.

[1] RDF allows for resources to be defined without URIs. These resources are called blank nodes. However, such resources do not appear in this thesis.

[2] http://www.w3.org/2004/02/skos/

[3] `https://en.wikipedia.org/wiki/Category:Semantic_Web`



- Latent semantic indexing
- Lattice Miner
- Library of Congress Linked Data Service
- Lightweight ontology
- Linked data
- Linked Data Platform
- LinkedGov

M

- Metaweb
- Microformat
- Minimal mappings
- Mulgara (software)
- Multimedia Web Ontology Language
- MultiNet

N

- Named graph

- Web of trust
- Web Ontology Language
- Web resource
- Web Rule Language
- Web Services Modeling Language
- WebID
- Wolfram Alpha
- WSMO
- WYSIWYM (interaction technique)

Y

- Yebol
- Yummly

Z

- Zemanta

Categories: Internet ages | World Wide Web Consortium | Knowledge representation | Data management

Hidden categories: Commons category with local link same as on Wikidata

Figure 2.1: A screenshot of the Wikipedia category page for ‘Semantic Web’.

Term Frequency

In the term frequency method the weight of a term indicates how frequently that term was mentioned relative to the other terms. An example of this can be seen in Table 2.4. In this table, a term's weight is calculated by dividing its number of mentions by the total number of mentions.

Term	**Frequency**	**Weight**
this	1	0.25
is	1	0.25
an	1	0.25
example	1	0.25

Table 2.4: An example document with term frequency-based weights

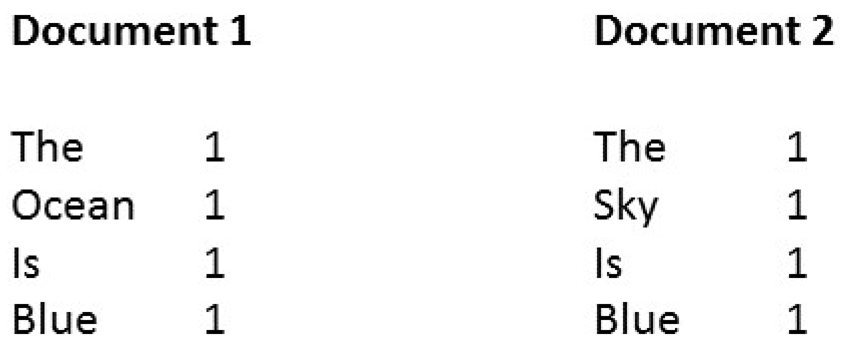
Hauff and Houben employ the term frequency-based representation in attempting to approximate a user's bookmarking profile using their tweets [31]. Gao et al. employ the term frequency-based representation in comparing user behaviour on Twitter with a similar Chinese site, Sina Weibo [46].

Term Frequency-Inverse Document Frequency

Term Frequency-Inverse Document Frequency (tf-idf) is a weighting scheme that assigns a weight to each term in a document based on the term's importance to the document collection as a whole [47]. It combines two measures: (i) Term Frequency - The number of times a term appears in a particular document (ii) Inverse Document Frequency - An inverse function of the number of documents in which the term appears. The weight $\mathrm{w}_{\mathrm{i,j}}$ of term *i* for document *j* is given by the following formula:

$$w_{i,j} = f_{i,j} \cdot \log_{10} \frac{N}{n_i} \tag{2.1}$$

In the above formula $\mathrm{f}_{\mathrm{i,j}}$ refers to the frequency of term *i* in document *j*, N refers to the total number of documents and n_{i} refers to the total number of documents in which term *i* appears. This formula accounts for the fact that certain terms appear often in many documents, and thus the fact that a term appears frequently in one document does not necessarily mean it is uniquely relevant for that document. Examples of terms that will often receive low tf-idf weights are stop words such as 'the', 'a' etc.



Document 1		Document 2	
The	1	The	1
Ocean	1	Sky	1
Is	1	Is	1
Blue	1	Blue	1

Figure 2.2: Two example documents with term frequencies.

Consider the example in Figure 2.2, showing two documents with their term frequencies. The weight of the word ‘blue’ for Document 1 is given by:

$$1 \cdot \log_{10} \frac{2}{2} = 0 \tag{2.2}$$

Since the word ‘blue’ appears in both documents, it does not have any special relevance for Document 1. On the other hand, the weight of the word ‘ocean’ for Document 1 is given by:

$$1 \cdot \log_{10} \frac{2}{1} = 0.301 \tag{2.3}$$

Gilbert et al. apply tf-idf to compute the similarity between Facebook users’ ‘interests’ and ‘about’ information in order to estimate the strength of the relationship between users [48]. Abel et al. apply tf-idf to generate rankings of users’ interests from their tweets for the purposes of news recommendation [49].

2.5 Part-Of-Speech tagging and Chunking

Part-Of-Speech (POS) tagging is the process of labelling the words in a text with their part-of-speech, e.g. noun, verb, adjective etc. POS-tagging forms a fundamental part of studies such as [36], [46] and [40], as it provides the linguistic information required to link words in text with information in external linguistic resources. POS-tagging has also been employed in studies such as Ding and Jiang [50] and Lioma and Ounis [51] in order to generate Bag-Of-Words representations. Ding and Jiang use POS-tags to identify words representing interests in a users’ Twitter biographies[4]. Lioma and Ounis use POS-tags to identify “content-rich”

[4] The biography section of a user’s Twitter profile contains a textual self-description by the user. This could be, for example, ‘John Smith, avid ornithologist, Game of Thrones fanatic’.

word sequences in search queries. Content-rich word sequences are those that express a large amount of information in the query. For example 'cheap restaurants' is content-rich, whereas 'in the' is not. An example of POS-tagging of the sentence S: 'John really is a tall man.' can be seen in Table 2.5. This sentence has been tagged using the Pattern NLP library for Python [52]. This library employs the POS-tagging method described in [53]. The tags come from the Penn treebank II dataset[5].

Word	**POS-tag**	**Tag meaning**
John	NNP	Noun, proper singular
really	RB	Adverb
is	VBZ	3rd person singular present verb
a	DT	Determiner
tall	JJ	Adjective
man	NN	Noun, singular or mass

Table 2.5: A POS-tagged sentence.

The process of POS-tagging facilitates a process called *chunking*, in which words are grouped together according to the grammatical roles they play in a text. Table 2.6 shows the chunks derived from S. The chunking is also performed using the Pattern library, which employs the algorithm described in [54]. The label 'noun phrase' signifies that the chunk consists of a noun which may be modified in some way by other words. In this case, there are two noun phrases, one containing an unmodified noun 'John' and the other containing a noun 'man' modified by the adjective 'tall'. The label 'verb phrase' signifies that the chunk consists of a verb which may be modified in some way by other words. In this case the verb 'is' is modified by the adverb 'really'.

Chunk	**Chunk tag**	**Tag meaning**	**Role**
['John']	NP	Noun phrase	Subject
['really', 'is']	VP	Verb phrase	None
['a', 'tall', 'man']	NP	Noun phrase	Object

Table 2.6: A chunked sentence.

[5] Described in full detail here: http://www.clips.ua.ac.be/pages/mbsp-tags

2.6 A survey of personalisation approaches that apply structured representations of unstructured content

This section surveys personalisation approaches that apply structured representations of unstructured content. Recall that Brusilovsky defines personalised applications as those that “build a model of the individual user and apply it for adaptation to that user” [4]. Brusilovsky and Millán identify the following user characteristics to which personalised systems can adapt [55]:

- Knowledge - The user’s knowledge of the subject of the content being analysed by the personalisation approach. For example, an e-learning application may adapt content by providing beginners with more high-level material while providing advanced learners with more technical material.
- Interests - Subjects in which the user is interested. For example, a movie recommendation application may model a user’s preference for Action films, Comedies etc.
- Goals - The objective the user is trying to achieve within the personalisation application. For example, in a web search system, the user’s goal is to retrieve web pages that correspond to their query.
- Background - Background refers to all information related to the user’s experience outside of the personalisation application that is still relevant for that application. For example, an e-learning application may leverage a user’s profession in determining which courses to display to the user.
- Individual traits - Defined by Brusilovsky and Millán as “the aggregate name for user features that together define a user as an individual”. Examples are personality traits (e.g., introvert/extravert), cognitive styles (holist/serialist), cognitive factors (e.g., working memory capacity) and learning styles.

However, the survey in this section will not examine all of the above characteristics. Recall that this thesis is not concerned with investigating particular personalisation approaches. As such, user characteristics that relate to particular personalisation approaches are not in the scope of this thesis. The following characteristics are therefore not examined in this survey:

- Goals - The definition of the Goals characteristic depends on the personalisation approach. Per Brusilovsky and Millán: “it can be the goal of the work (in application systems), an immediate information need (in information access systems), or a learning goal (in educational systems).”

- Background - The purpose of the Background characteristic is to provide additional information about the user required by a particular personalisation approach. For example, certain approaches for adaptive presentation of medical content have defined Background as the user's level of medical experience. On the other hand, certain language teaching approaches have defined Background as the user's language ability (e.g. novice, intermediate etc.) [3].

Individual traits are also not considered in this survey. This is because these traits are derived through leveraging principles of psychology such as psychometrics. These principles are outside of the scope of this thesis.

Note that Brusilovsky and Millán also identify the user's context as important in the personalisation process. Context refers to the general environment in which the personalisation is provided. Context can be defined as location, physical environment (e.g. heat, light conditions), affective state etc. Context is also specific to particular personalisation approaches. For example, a route planner may define context as the user's current location, while a music recommendation system may define context as the user's social network.

This survey will focus on personalisation approaches that model the Interests and Knowledge characteristics. It is important to note the variety of personalisation applications covered by these characteristics. On Knowledge, Brusilovsky and Millán state: "The user's knowledge of the subject being taught or the domain represented in hyperspace appears to be the most important user feature, for existing AES (Adaptive Education Systems) and AHS (Adaptive Hypermedia Systems)." These authors further state that the Interests characteristic is the most important characteristic in information retrieval systems and filtering systems that handle large amounts of information, as well as recommender systems.

Recall from Chapter 1 that personalisation approaches such as Steichen et al. [5], Hecht et al. [6] and Abel et al. [7] use structured representations of unstructured content in one of three ways: (i) To represent content written by the user to whom the approach is adapting (ii) To represent content written by someone other than the user to whom the approach is adapting (iii) A combination of (i) and (ii). The survey in this section is thus divided into three sub-sections, each of which examines personalisation approaches for each of (i), (ii) and (iii). Within each of these sub-sections, personalisation approaches adapting to the Knowledge and Interests characteristics are examined. Each personalisation approach will be critiqued as to whether it reveals how the choice of structured representation affected personalisation performance.

2.6.1 Personalisation approaches representing content written by the user

This section examines personalisation approaches that use structured representations to represent unstructured content written by the user to whom the approach is adapting.

2.6.1.1 Knowledge

Hauff and Houben aim to use a user's tweets to approximate that user's profiles on the bookmarking sites Bibsonomy[6], CiteULike[7] and LibraryThing[8] [31]. These authors employ the Bag-Of-Words representation. The tweets are first preprocessed to convert words to regular English. For example, repeated vowel sequences of three or more are converted to two (e.g. 'goood' becomes 'good'). Stopwords are then removed. The remaining words are combined into a single list. Words are weighted according to the frequency with which they occur. Hauff and Houben find that these knowledge profiles suffer heavily from noise introduced by tweets not related to a user's learning activities. Hauff and Houben define the solution to this problem as automatically identifying the tweets that should be used in building a knowledge profile. These authors propose using style information (e.g. whether or not the tweet is a retweet, the number of letters and characters in the tweet) for this task. Hauff and Houben report that this approach yields "inconsistent" results. However, these authors do not consider the possibility of leveraging an external linguistic resource. Rather than using the words in tweets directly, these words could be enriched using an external linguistic resource such as DBpedia or WordNet. This would then represent the user's profile as a list of subjects mentioned in the user's tweets, rather than a list of words. In such a representation, it is possible that noise in the form of irrelevant words could be reduced. This could be for example because synonymous words would be mapped to the same subject. In the Bag-Of-Words representation, 'philosopher' and 'philosophical' are distinct terms. With the help of the conceptual hierarchy in an external linguistic resource, each of these words could be mapped to the subject 'Philosophy'.

Stan et al. describe an approach for learning a user's levels of interactivity with, and expertise in, various topics from the content that user posts on various social media websites [56]. These authors employ the Named Entity-based representation. First, AlchemyAPI[9] is used to identify keywords in the social media content. These keywords are then linked to Named Entities in DBpedia by exploiting contextual clues in the user's previous conversations. For

[6] `https://www.bibsonomy.org/`

[7] `http://www.citeulike.org/`

[8] `https://www.librarything.com/`

[9] `http://www.ibm.com/watson/alchemy-api.html`

example, in the sentence 'I like apple', 'apple' can have multiple meanings. However, if the user has also posted about the IPhone and the Mac, it can be inferred that 'apple' refers to the technology company. Stan et al. exploit the hierarchical links in DBpedia to categorise the identified Named Entities. These authors then generate scores indicating the user's level of interactivity and expertise with respect to the identified concepts. However, Stan et al.'s approach does not consider the possibility that a structured representation consisting of categorised Named Entities may be lacking contextual information that was referenced indirectly in user posts.

2.6.1.2 Interests

Michelson and Macskassy and Kapanipathi et al. aim to perform the same task: identifying a user's interests from their tweets [21] [22]. Both approaches are similar in that they identify Named Entities in tweets and then categorise these Named Entities. However, each approach adopts a different process for identifying Named Entities and generating categories. Michelson and Macskassy identify capitalised words as candidate Named Entities. These authors then use Wikipedia to identify the entities to which these candidates refer. For each candidate, the Wikipedia articles to which the candidate can potentially refer are obtained. Then, the text of the tweet containing the entity is compared with the text of each candidate article. The article that has the most words in common with the tweet is chosen as the correct article for the entity. Once the Named Entities have been identified, categories are obtained by exploiting the Wikipedia category hierarchy. For example, if 'Theo Walcott' is identified as a Named Entity, 'English football players' is identified as a category. Kapanipathi et al. use Zemanta to identify Named Entities[10]. To obtain categories, these authors first process Wikipedia to remove cycles in its category relationships. Kapanipathi et al. then exploit hierarchical links in the processed version of Wikipedia to identify categories for Named Entities. In both of these approaches, the goal is to obtain latent categories of user interest. However, both approaches assume that Named Entities alone are a sufficient basis for obtaining these latent categories. If this is not the case, then certain latent categories may be missed. Each approach is evaluated solely by determining the accuracy with which information were identified. It is therefore unclear whether the Named Entities are a sufficient basis for capturing user interests.

Hecht et al. aim to determine a user's interests by analysing questions the user posts on Facebook [6]. These authors aim to leverage the user's social network to provide answers to these questions. Hecht et al. identify Named Entities by designing three different Named

[10] The Zemanta API no longer appears to be available. The link provided in approaches such as Kapanipathi et al. [22] and Orlandi et al. [29], `http://developer.zemanta.com/` is broken.

Entity Extractors. Identified Named Entities are used as a basis for searching the user's social network for relevant friends. For example, if a location is identified, friends with this location in their Facebook location field will be identified. If no Named Entities are identified in a question, that question is not answered. Hecht et al. report various questions to which their approach provides answers. However, these authors also report that their approach gave answers for 26.7% of all questions that were asked. This means that the majority of questions went unanswered. This in turn means that there is information in a large number of user questions that is not captured by the Named Entity-based representation. However, this is not considered in Hecht et al.'s approach. It is therefore unclear whether an approach that focuses on Named Entities is sufficient for this task.

Gao et al. describe a user modeling system called GeniUs that processes messages a user has written in order to create user models [57]. It does this by identifying Named Entities and using these Named Entities to derive concepts. Named Entities are identified using DBpedia Spotlight [58], which links text to items in DBpedia. Concepts are identified using DBpedia relations. The contents of the user models created by GeniUs can be filtered according to application requirements. For example, a book recommender can create models containing only book-related terms. The weights of items in the user model can also be modified according to application requirements. For example, a term's weight can be assigned based on the recency with which the term was mentioned. Gao et al. evaluate their system in the context of a tweet recommendation experiment. Participants' Twitter posts (tweets) are analysed in order to create a complete Twitter-based profile as well as domain-specific profiles. The domain-specific profiles contain only subjects related to a specific domain e.g. products such as books, software etc. Gao et al. find that the domain-specific profiles provide better tweet recommendations than the complete Twitter profile. This finding demonstrates that adapting a structured representation with respect to content domain yields an improvement in personalisation performance. However, this evaluation does not address how the underlying representation affected recommendation performance. In other words, Gao et al.'s evaluation informs about different variants of their chosen representation, but not about that representation itself.

Certain approaches aim to aggregate user information from multiple sources in order to combine user interests and knowledge together into a single user model.

Abel et al. describe a system called Mypes that allows users to aggregate their different social media profiles [28]. These authors aim to find commonalities between tags in different tagging services by mapping these tags to WordNet synsets. For example, the ‘research’ tag is mapped to ‘cognition’, the ‘hypertext’ tag is mapped to ‘communication’. The mapped tags from the individual services are then combined to create an aggregated tagging profile. These authors evaluate their approach in the context of a tag prediction task. Abel et al. compare an aggregation and non-aggregation strategy in this task and find that the aggregation strategy outperforms the non-aggregation strategy. These authors also find that enriching tags with WordNet synsets improves recommendation performance. This second finding highlights the benefits of applying the conceptual hierarchy contained in an external linguistic resource in describing user information from multiple sources. Mapping information in different profiles to a single representation (in this case, WordNet synsets) facilitates the task of profile aggregation. However, as no other external resource is employed, it is unclear whether WordNet is the most suitable resource for this task.

Orlandi et al. also aim to create aggregated user profiles [29]. These authors leverage DBpedia for the task. Orlandi et al. identify Named Entities in user social media posts using Zemanta. Once Named Entities have been identified, associated concepts are identified using links in DBpedia. The number of occurrences of each entity is recorded, as is the timestamp of the post in which the entity was identified. Orlandi et al. use this information to create a weight that reflects both the regularity and frequency of occurrence of each entity. This approach yields separate sets of Named Entities for each social media service, which are then combined to create an aggregated set. These authors evaluate their approach by asking participants to evaluate the accuracy of their aggregated profile. Orlandi et al. compare their approach against a Bag-Of-Words representation using term frequency weights alone. These authors find that their approach substantially outperforms the Bag-Of-Words approach. Orlandi et al. also ask participants to list the entities and concepts that describe their information. These authors then compare the participants’ aggregated profiles with this list to assess the extent to which the profiles cover the user information. However, as these authors themselves point out, this method is flawed since participants cannot necessarily be expected to provide a comprehensive list of entities and concepts to describe their information. Furthermore, since only DBpedia is used, there is no basis for comparing the coverage of these authors’ approach.

2.6.2 Personalisation approaches representing content not written by the user

This section examines personalisation approaches that use structured representations to represent unstructured content written by someone other than the user to whom the approach is adapting.

2.6.2.1 Knowledge

Heitmann et al. aim to model user expertise across multiple domains in an enterprise environment [37]. These authors represent user expertise and items to be recommended as lists of DBpedia resources. Items are recommended to users based on their conceptual similarity to users' interests. Heitmann et al. describe their approach in the context of a use case in which employees in an organisation are aided in finding content that is relevant to their area of expertise. However, it is not necessarily the case that these authors' chosen representation will always facilitate this recommendation. It may the case that this kind of representation will not facilitate effective recommendation in certain domains. This possibility is not examined.

Guo et al. compare three different approaches for recommending jobs [30]:

- Content-based - This involves using a description of an item to determine whether the item should be recommended. Guo et al. use unstructured job descriptions.
- Case-based - This involves using structured information about a job e.g. title, level of experience required etc. to determine whether an item should be recommended.
- Hybrid - A combination of content-based and case-based.

Guo et al. employ a variety of structured representations for the unstructured job descriptions, using Bag-Of-Words, Named Entities and Wikipedia categories. These authors employ OpenCalais[11] to obtain Named Entities and Wikipedia categories. Guo et al. also compare tf-idf weighted, document frequency weighted and non-weight approaches. Guo et al. find that the Bag-Of-Words representation performs best overall. However, the authors do not fully address the reasons for this result. This result appears to suggest that the Bag-Of-Words representation captured important aspects of the jobs that the other structured representations did not. However, it is not clear what aspect this is.

2.6.2.2 Interests

Semeraro et al. describe an approach for recommending documents to users based on documents those users have previously read [40]. Each document is represented as a list of WordNet synsets that have been identified in the text of the document. Each synset is weighted according to the number of times that synset occurs. Semeraro et al. adopt three different methods for

[11] http://www.opencalais.com/opencalais-api/

linking words in text to WordNet synsets: one for adverbs and adjectives, one for nouns and one for verbs. A user's user model consists of the synsets associated with the documents they have read. New documents are recommended to the user based on the similarity between those documents and the user model. Semeraro et al. evaluate their approach in the context of a system that recommends academic conference talks. These authors compare their approach with an approach that applies the Bag-Of-Words representation[12]. Semeraro et al. find that their approach provides more accurate recommendations than the Bag-Of-Words approach. This demonstrates the utility of leveraging external linguistic resources in recommending content to users. However, since the authors' approach is only applied in a single domain, it is unclear whether this approach is generally effective for document recommendation, or only in this domain. Furthermore, these authors do not compare WordNet with another linguistic resource, making it unclear whether WordNet is the most suitable resource for this kind of task.

Mendes et al. employ the Named Entity-based representation to provide a service that allows users to query Twitter for tweets related to particular news items [59]. In Mendes et al.'s system, users subscribe to 'Concept Feeds'. These feeds consist of collections of tweets identified according to user queries. These queries select specific subsets of tweets based on criteria such as the time a tweet was posted, the location of the tweet's author or the specific properties of the Named Entities the tweet contains. Named Entities are identified by means of dictionary lookup using a list of DBpedia entities. As new tweets are published to a Concept feed to which a user has subscribed, these tweets are delivered to the user in real-time. Mendes et al. argue that their approach allows users to make sense of the large amount of data on the web by allowing them to focus only on the parts of interest to them. However, the authors do not address the extent to which the Named Entity-based representation facilitates this task. For example, consider the example tweet in Figure 2.3, taken from Mendes et al. The authors' system extracts 'Health-Care Reform' and 'Mitt Romney' from this tweet. However, the information provided by the text 'Left Wing' is also relevant for this tweet.

```
Left Wing Health-Care Reform, Mitt Romney, and Next Steps
```

Figure 2.3: An example tweet from Mendes et al.

[12] Described in Section 2.2.

2.6.3 Personalisation approaches representing content written by the user and content not written by the user

This section examines personalisation approaches that use structured representations to represent both content written by the user and content not written by the user to whom the personalisation approach is adapting.

2.6.3.1 Knowledge

Heap et al. describe an approach that leverages a user's career path in order to recommend jobs for that user [60]. These authors use the Bag-Of-Words representation to represent users and jobs to be recommended. Heap et al. use the user's LinkedIn profile in order to obtain information about the user's career. The user's previous positions are extracted directly. Nouns and noun phrases are extracted from the user's long-form information using the OpenNLP library[13]. These authors compare their approach with a baseline approach that does not consider career transition. Heap et al. find that the baseline approach outperforms their approach in recommending jobs. However, these authors do not address their choice of structured representation. Recall that Orlandi et al. find that the Named Entity-based representation outperformed the Bag-Of-Words representation in representing user interests. It is therefore possible that by employing a different representation, Heap et al. may have been able to achieve better recommendation performance.

Piao and Breslin aim to leverage a user's LinkedIn profile to recommend Massive Open Online Courses (MOOCs) [61]. These authors generate separate user models for each of the user's job titles, educational fields and skills in order to determine which user model provides the best recommendations. User models and MOOCs to be recommended are represented using the Bag-Of-Words representation. Piao et al. find that the user model consisting of skills yields the best MOOC recommendations. However, the reasons for this result are not clear. Also, since no other structured representation is employed, it is not clear whether this result holds in general or just for the Bag-Of-Words representation.

2.6.3.2 Interests

Li et al. aim to leverage users' posts on Sina Weibo (a Chinese microblogging site) to recommend books to those users [62]. These authors employ the Bag-Of-Words representation to represent user interests and books. Li et al. group books into clusters e.g. an Action

[13] `https://opennlp.apache.org/`

books cluster, a War books cluster etc. The clusters are ranked according to their similarity with the user model, and books are recommended from the top-ranked clusters. Li et al. compare their approach against a number of baselines, including recommending the most popular books and recommending books at random. These authors find that their approach outperforms these baselines. However, Li et al. do not examine the relationship between the Bag-Of-Words representation and their recommendation results. These authors report an accuracy of approximately 27% for their approach, indicating a substantial amount of error. The extent to which the choice of structured representation contributed to this error is not examined.

Ko et al. aim to recommend new items to a user based on their friends' interests [63]. These authors analyse Facebook data to perform this task. Interests are represented as concepts using three different ontologies, DBpedia, Freebase[14] and OpenCyc[15]. These authors extract nouns and noun phrases and link these with concepts using the SPARQL query language[16]. Similar concepts are grouped together, thereby representing the user's information as a collection of semantic clusters. The user's friends' information is represented in the same way. Similarity is then computed between the user's semantic clusters and their friends' semantic clusters by identifying shared concepts. This gives a ranking of the friends' concepts on the basis of similarity. New concepts are then recommended for the user from the top-ranked clusters. Ko et al. evaluate their approach by means of a user study. Participants are asked to indicate their satisfaction with concepts that were recommended for them. These authors' chosen baseline is a previous research approach for recommending interests on Facebook. Ko et al. find that their approach outperforms this baseline. However, these authors' approach employs only general knowledge bases and does not consider lexical resources such as WordNet. As such, it is difficult to determine the overall effectiveness of these authors' approach.

2.6.4 The characteristics of structured representations of unstructured content

From the preceding discussion on personalisation approaches applying structured representations of unstructured content, the following five characteristics emerge:

1. The type of information contained in the structured representation e.g. Named Entities in Hecht et al. [6], concepts in Ko et al [63].

[14] `https://developers.google.com/freebase/` (Deprecated)

[15] `http://sw.opencyc.org/`

[16] SPARQL allows applications to request information from an RDF database. The language specification can be found here: `https://www.w3.org/TR/rdf-sparql-query/`

2. The external resource employed to generate the structured representation e.g. WordNet in Abel et al. [28], DBpedia in Orlandi et al. [29].

3. The domain of the unstructured content being analysed, e.g. jobs in Heap et al. [60], news in Mendes et al. [59]

4. The analysis method employed to generate the structured representation e.g. dictionary lookup in Mendes et al. [59], DBpedia Spotlight in Gao et al. [57]

5. The weighting method employed e.g. term frequency in Hauff and Houben [31], frequency and regularity of occurrence in Orlandi et al. [29]

Note that characteristics 4 and 5 are approach-specific. For example, each of Michelson and Macskassy and Kapanipathi et al. apply the Named Entity representation to identify a user's interests from their tweets. However, the way in which the Named Entity representation is generated differs for each approach. Similarly, the way in which terms are weighted reflects the emphasis an approach wishes to put on certain terms. For example, Gao et al.'s GeniUs system allows for terms to be weighted as a function of time. Therefore, characteristics 4 and 5 are not the focus of this thesis.

The preceding discussion also highlighted gaps with respect to characteristics 1,2 and 3. These gaps are summarised in the following paragraphs.

Characteristic 1 - The type of information contained in the structured representation

Hecht et al. employ the Named Entity representation in their approach for implementing personalised search [6]. However, these authors' evaluation does not examine the kinds of queries that cannot be answered using this representation. Michelson and Macskassy [21] and Kapanipathi et al. [22] identify Named Entities in tweets and categorise these Named Entities to obtain concepts. However, these concepts are in fact categories of Named Entities, which may not fully capture the conceptual information contained in the tweets. For example, contextual information referenced indirectly in the tweets is not captured. Approaches such as these do not give a clear view as to what content-based representations such as Named Entities can and cannot represent about users.

Characteristic 2 - The external resource employed to generate the structured representation

Both Abel et al. [28] and Orlandi et al. [29] leverage the conceptual hierarchy contained in an external linguistic resource to combine user information from multiple services in order to represent multiple user characteristics. However, the former employ WordNet while the latter employ DBpedia. Recall from Chapter 1 that each of these resources is organised according to different principles. This means that each resource will yield different representations of user information. This fact warrants a comparison of these resources in order to investigate which provides the most accurate representation of user information.

Characteristic 3 - The domain of the unstructured content being analysed
Recall that Guo et al. find that the Bag-Of-Words representation yields better job recommendations than structured representations consisting of concepts and Named Entities. However, these authors' evaluation does not reveal what property of the jobs that was captured by the Bag-Of-Words representation but not captured by the other representations. Recall also that Heitmann et al. aim to represent content of different domains. However, these authors do not account for the fact that their chosen representation may not accurately represent every domain.

2.6.5 Required properties of the experiments in this thesis

Recall from Section 1.2 that the research method of this thesis consists of in-depth investigations of individual representative cases of general phenomena. Taken together, these cases must address a variety of personalisation tasks, as well as treating content of multiple formats and domains. However, each case must have its own individual properties that allow for a full investigation of the individual characteristic it repreresents. These individual properties are described in this section.

Characteristic 1 - The type of information contained in the structured representation
This experiment will address the capabilities and limitations of the Named Entity and Bag-Of-Words representations. The experiment will analyse content users have created in order to examine the different ways in which these users express themselves. The experiment must investigate both content- and context-dependent forms of user expression, in order to determine the ability of Named Entities and Bag-Of-Words to represent both. The purpose of this experiment is to clearly identify how the inability of the Named Entity and Bag-Of-Words representations to capture context limits their ability to fully describe different forms of user expression. This requires that the content being analysed be of a single, narrowly defined domain. If content from multiple domains were analysed, the ability of each representation to accurately represent each domain would also be a factor in determining is ability to represent

different forms of user expression. This could potentially obscure the analysis into the effect of the lack of context in both representations.

Characteristic 2 - The external linguistic resource employed to generate the structured representation

This experiment will investigate the ability of different external linguistic resources to provide the context that is absent in content-based representations such as Named Entities and Bag Of Words. This experiment will compare the efficacy of different linguistic resources in representing user interests and knowledge. This experiment will investigate the relationship between a linguistic resource's social relevance and its ability to represent these characteristics. This experiment must treat user information from multiple profiles. This is because users distribute their information across different profiles [28] [29]. Each linguistic resource must therefore be able to accurately represent user information from multiple profiles. This experiment must assess the ability of each resource to represent information from these profiles separately, as well as to combine them.

Characteristic 3 - The domain of the unstructured content being analysed

This experiment will investigate the way in which content domain affects the efficacy of structured representations in personalisation tasks. This experiment will examine the task of content recommendation across different content domains, user models and recommendation methods. This experiment will investigate how a structured representation's ability to reflect domain category distinctions affects recommendation performance irrespective of domain, user model or recommendation approach.

2.7 Conclusion

This chapter discussed methods for generating structured representations of unstructured content for the purposes of personalisation. This discussion examined different types of representation, different methods for generating weights and different kinds of external resources that can be employed. The Part-Of-Speech tagging and chunking methods were also discussed. The final section of this chapter highlighted three gaps with regard to the choice of structured representations for personalisation tasks. The first of these gaps will be addressed in the next chapter.

Chapter 3

Examining the Named Entity-based representation

The first experiment in this thesis investigates the Named Entity and Bag-Of-Words representations. This investigation concerns the inability of Named Entities and Bag-Of-Words to capture context, and how this affects the ability of these representations to represent the different ways in which users express themselves. This experiment is described in this chapter.

Section 3.1 describes the objective of this experiment. **Section 3.2** describes the case that will be investigated in this experiment. **Section 3.3** describes the Twitter microblogging service. **Section 3.4** describes the question of this experiment. **Section 3.5** describes the process for generating the structured representation used in this experiment. **Section 3.6** describes the experiment design. **Section 3.7** discusses the results of the experiment and **Section 3.8** concludes.

3.1 Objective of this experiment

Recall from Section 1.1 that the first objective of the research question of this thesis is to investigate how the inability of Named Entities and Bag-Of-Words to capture context limits the ability of these representations to fully represent different forms of user expression. Recall also the requirements for this experiment described in Section 2.6.5. The objective of this experiment is thus to investigate whether the inability of these representations to capture context prevents them from representing: (i) Certain forms of user expression entirely (ii) Variants of certain forms of user expression (iii) A combination of (i) and (ii). This experiment will treat a single narrowly defined domain. This ensures that the analysis of the absence of context in these representations is not obscured by the ability of each representation to accurately represent multiple domains.

3.2 Case description

Social media services are used widely by users to express themselves and engage with other users [64]–[66]. One of the more popular social media services is Twitter, which counts 338 million montly active users [67]. The author thus argues that Twitter is a representative example of social media services.

Since this experiment must treat a single domain, it is necessary to filter user tweets so as to only obtain tweets relating to a single domain. One possible approach to this task is to obtain tweets containing a specific hashtag. However, not all tweets contain a hashtag. In fact, in a study of 62 million tweets, Hong et al. find that the percentage of tweets that contain a hashtag is 11% [68]. Thus, a case study focusing only on tweets that contain a particular hashtag would not necessarily yield generalisable results. Instead, tweets relating to a single domain can be obtained by obtaining the tweets sent to domain-specific Twitter accounts. This is not guaranteed by taking general organisational or individual accounts. This is because tweets to these accounts could reference any aspect of the associated person or organisation. However, customer support Twitter accounts meet this requirement. These accounts have been created solely to receive customer support-related tweets from users. In the experiment described in this chapter, the BusinessDictionary definition of customer support is used. It is as follows: "Range of services provided to assist customers in making cost effective and correct use of a product" [69]. BusinessDictionary is chosen as a reference as it has been employed in various research approaches. Diirr et al. use the BusinessDictionary definition of 'process' as a basis for developing a tool that facilitates discussion between Government and Society about public service processes [70]. Weigand et al. apply the BusinessDictionary definition of 'policy' in creating a framework that allows enterprises to automate the management of their services [71]. Van Hise et al. reference the BusinessDictionary definition of 'professionalism' in their description of a method for designing an accounting ethics course [72].

Naaman et al. [73] and Honeycutt and Herring [74] define various ways in which users express themselves on Twitter. These are given in tables 3.1 and 3.2 respectively:

Information Sharing e.g. “15 Uses of WordPress”	**Self Promotion** e.g. “Check out my blog I updated 2day 2 learn abt tuna!”	**Opinions / Complaints** e.g. “Go Aussie $ go!”	**Statements and Random Thoughts** e.g. “The sky is blue in the winter here”	**Me now** e.g. “tired and upset”
Question to followers	**Presence Maintenance** e.g. “i’m backkkk!”	**Anecdote (me)**	**Anecdote (others)**	

Table 3.1: Naaman et al.’s forms of user expression

About addressee: solicits or comments on information relating to the addressee	**Announce/advertise**: announces information to the general readership of Twitter	**Exhort**: directs or encourages other(s) to do something	**Information for others**: posts information appar- ently intended for others; may be solicited or volunteered
Information for self: posts information apparently intended for sender’s own use	**Metacommentary**: comments on Twitter or twittering	**Media use**: reports or reflects on media use, especially music	**Opinion**: asserts a subjective or evaluative position
Other’s experience: solicits, reports on, or comments on information relating to the experience of a third person or persons	**Self experience**: reports or comments on sender’s own experience	**Solicit information**: requests information (other than about addressee)	**Other**: miscellaneous other themes, e.g., greetings, nonsense

Table 3.2: Honeycutt and Herring’s forms of user expression

There is a large degree of overlap between these different categories, shown in Table 3.3.

Naaman et al.	**Honeycutt and Herring**
Information sharing	Announce/advertise, Information for others
Opinions/Complaints	Opinion
Statements and Random Thoughts, Presence Maintenance	Other
Me now	Information for self
Question to followers	Solicit information
Anecdote (me)	Self experience
Anecdote (others)	Other's experience

Table 3.3: Overlapping categories between Naaman et al. and Honeycutt and Herring.

Taking these overlapping categories and categories from each approach that have no overlap yields 13 categories. However, since metacommentary is specific to Twitter, findings related to this category would likely not be generalisable. This category is therefore removed, giveing a total of 12 categories, shown in Table 3.4.

Information sharing	Opinions/Complaints	Statements and Random Thoughts	Me now
Question to followers	Presence Maintenance	Anecdote (me)	Anecdote (others)
Self promotion	Exhort	About addressee	Media use

Table 3.4: The combined list of categories from Naaman et al. and Honeycutt and Herring.

Note that these forms of expression are not specific to Twitter. For example, the OECD lists behaviours such as information sharing, expressing opinions and marketing (i.e. exhorting individuals to buy a product) as being associated with social media services in general [75].

This chapter describes a content analysis performed on 38976 English-language tweets sent to or by the customer support Twitter accounts of distinct organisations in the US, Europe and Asia. Tweets sent to these customer support accounts are analysed in order to represent the different forms of user expression given in Table 3.4.

The author argues that this case meets the requirements of Experiment 1 described in Section 2.6.5 as the categories of self-expression in Table 3.4 comprise are a mixture of content- and context-dependent. For example, one tweet Naaman et al. provide as an example of 'Information Sharing' is "15 Uses of WordPress". In this case, the interpretation of this tweet derives directly from the tweet text itself. However, consider one tweet provided by Naaman et al. as an example of 'Opinions/Complaints', "Go Aussie $ go!". This sentence refers to the Australian dollar, but references an external context e.g. the Australian dollar's exchange rate.

3.3 Research Question of this Experiment

The research question of this experiment is as follows:

How does the absence of context in content-based representations limit their ability to represent different forms of user expression?

This question will be examined with respect to each of the forms of user expression given in Table 3.4. This will give two perspectives as to the limitations of content-based representations such as Named Entities and Bag-Of-Words:

1. Are there categories of user expression that Named Entities and Bag-Of-Words cannot represent? For example, are they unable to represent the 'Opinion' category?

2. Are there categories of user expression that Named Entities and Bag-Of-Words can only partially represent? For example, are they able to represent some instances of the 'Information Sharing' category but not others?

This will allow for the limitations of Named Entities and Bag-Of-Words to be examined within and between the different categories of user expression. Thus, the question of this experiement provides a full investigation of the objective described in Section 3.1.

3.4 Generating the structured representation

Named Entity and Bag Of Words representations are created for each customer support account. Named Entities and Bag-Of-Words for each customer support account are weighted using tf-idf, where the tf term is the number of times the term appeared in a customer support account and the idf term is the number of times the term appeared in all other accounts combined[1]. The following subsections describe the process used to generate the Named Entity and Bag-Of-Words representations employed in the experiment described in this chapter.

[1] For more information about tf-idf, please see page 18 of this thesis.

3.4.1 Named Entities

Named Entities are identified using the AlchemyAPI Entity Extraction tool[2]. This decision was made with reference to studies such as [76] and [77], in which AlchemyAPI is employed in extracting information from tweets. Furthermore, Rizzo et al. describe a comparison between AlchemyAPI and four other popular Named Entity extractors and find that AlchemyAPI achieves the best entity extraction performance [78]. The AlchemyAPI Named Entity Extractor extracts links words in text with entities in various external resources. The experiment described in this chapter makes use of the DBpedia knowledge base, which is described in Section 2.3. The decision to use DBpedia is motivated by the fact that its information is derived from Wikipedia, the quality of whose information almost matches that of Encyclopedia Britannica [79].

3.4.2 Bag Of Words

The Bag Of Words representation is generated using the TfidfVectorizer[3] package provided by the scikit-learn library[4]. This package automatically creates a tf-idf-weighted Bag Of Words representation from text.

Abel et al. argue that tweets can be further enriched by also analysing linked webpages [7]. However, this poses the following problems for the experiment described in this chapter:

- There is no guarantee that the linked document relates specifically to customer support, as in Figure 3.1. The link in this tweet went to the website for a San Francisco radio station.
- There is no guarantee that links are not spam, as in Figure 3.2. The link in this tweet went to a short term loan site[5].

Analysing links such as the two in Figures 3.1 and 3.2 would skew the results of this study. For example, entities identified in the first link would relate to shows the station is airing, which are

[2] http://www.alchemyapi.com/

[3] `http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html`

[4] `http://scikit-learn.org/stable/`

[5] The link for this site is now broken. A Google search for this link yielded the following site: `http://homeinsurances.org/`

```
@comcastcares Why does Comcast advertise on Limbaugh on
stations like http://t.co/x5BzL4Vg?  Do you condone his
statements on Sandra Fluke?
```

Figure 3.1: A tweet containing a link un-related to customer support.

```
Keep your children, pets and parents safe on one site:  #RT
http://t.co/DwF8ueRs @comcastcares @designtaxi @bill_starr
```

Figure 3.2: A spam tweet.

completely unrelated to customer support. Thus, in the experiment described in this chapter, Named Entity and Bag-Of-Words representations are created using only tweet text.

3.5 Experiment design

3.5.1 Dataset Description

Over an 11-week period from 23rd February to 10th May 2012 (weekends excluded) 38,976 tweets were collected using the Twitter REST API. This number is significantly smaller than that in [80], despite the collection period in this experiment being over twice as long. This suggests that, for these organisations, Twitter is not a heavily used form of customer support.

3.5.1.1 Accounts Chosen

As stated above, organisations headquartered in various different regions were chosen. Companies were chosen from a variety of industries. These industries are as follows:

- Telecommunications
- Broadcasting
- Computer Software and Hardware
- Information Technology
- Printer Manufacturing
- Communications
- Electronics

@ATTCustomerCare (AT&T, U.S.A)	@comcastcares (Comcast, U.S.A.)	@dellcares (Dell, U.S.A)	@eircom (Eircom, EU)
@HPSupport (Hewlett-Packard, U.S.A)	@imagineWiMax (Imagine, EU)	@LexmarkListens (Lexmark, U.S.A)	@MicrosoftHelps (Microsoft, U.S.A)
@NokiaHelps (Nokia, EU)	@NortonSupport (Symantec, U.S.A)	@O2IRL (O2 Ireland, EU)	@OrangeHelpers (Orange, EU)
@SamsungSupport (Samsung, Asia)	@Sonylistens (Sony Electronics, Asia)	@SupportApple (Apple, U.S.A)	@TMobileUKhelp (T-Mobile UK, EU)
@UPC_HelpsYou_IE (UPC, EU)	@VZWSupport (Verizon Wireless, U.S.A)	@YahooCare (Yahoo!, U.S.A)	@VodafoneUK (Vodafone UK, EU)

Table 3.5: Customer support accounts chosen

In total, 20 customer support accounts were selected. These are listed in Table 3.5. Some example tweets can be seen in Figures 3.3 and 3.4.

```
@ATTCustomerCare I have been experiencing frequent issues
still with my dsl service (dropping, outages).  Already had
someone out.  Help?
```

Figure 3.3: A tweet to the AT&T account.

```
@MicrosoftHelps that my account is temporarily blocked, and
all of the links to help, and bring me back to that same
message.
```

Figure 3.4: A tweet to the Microsoft account.

3.5.2 Evaluation Method

Each of the Named Entity and Bag-Of-Words representations is assessed using the following steps:

1. For each account, take identified terms (either Named Entities or Bag-Of-Words) and their associated tweets.

2. Categorise these tweets according to the forms of user expression given in Table 3.4.

3. Determine whether each representation has either fully or partially captured the subject of the tweet. For example, is the mentioned word/entity the sole subject of the tweet, or does it reference a broader context of which the entity/word is a part?

3.5.3 Results

The top ten mentioned entity types, with their relative mention scores, are given in Table 3.6. As can be seen from the table, organisation and location entities constitute the vast majority of entity mentions. Indeed, the first entity of another type appears at position 7 in the table. Furthermore, the cumulative mention scores for the entities in positions 1-6 is 90.94%. The remaining four types in the table (combined with the remaining sixty that are not shown) thus account for less than 10% of Named Entities mentioned. For this reason, the the tweets discussed in this section contain primarily these two types.

Entity Type	**Relative Score**
Organisation	27.60
Company	26.18
Place	12.82
Populated Place	12.73
Country	9.80
Settlement	1.81
Person	1.24
Broadcaster	0.88
City	0.77
Administrative Region	0.77

Table 3.6: The top ten entity types, ranked by relative mention.

For consistency, the Bag-Of-Words representation must also be reduced. However, the Bag-Of-Words representation has no similar type information that can be used to remove terms with few mentions. Therefore, the Bag-Of-Words lists for each account are reduced using cumulative tf-idf weights. That is, only the terms with a cumulative tf-idf weight of 91% are retained.

There are 12 subsections in this section; one for each of the forms of user expression given in Table 3.4. Each subsection of this section will first provide examples of tweets where the Named Entity and Bag-Of-Words representations were sufficient for interpreting the subject of the tweet. Each subsection will also provide examples of tweets where these representations were insufficient for interpreting the subject of the tweet.

3.5.3.1 Information Sharing

Where Named Entities and Bag-Of-Words are sufficient

Consider the tweets in Figures 3.5 and 3.6, sent to the Microsoft and TMobile UK accounts respectively. In 3.5, a user informs other users that they can avail of customer support from Microsoft over Twitter. In 3.6, a user informs other users of an upcoming update for a mobile phone.

`Cool! Just found out that you can get #Microsoft support over twitter by tweeting to @MicrosoftHelps.`

Figure 3.5: An information sharing tweet to the Microsoft account.

`Galaxy S II ICS update available in UK starting March 19th, subject to carrier testing http://t.co/VaGoB9Ty look at. This @TMobileUKhelp`

Figure 3.6: A tweet to the TMobile UK account.

Where Named Entities and Bag-Of-Words are insufficient

Consider the tweet in Figure 3.7, to the Eircom account. In this tweet, a user appears to be criticising Eircom for blocking the Pirate Bay file sharing service. The Named Entitiy representation identifies 'Virgin Media' and the Bag-Of-Words representations identifies 'Virgin' and 'Media' for this tweet. However, both representations fail to represent the fact the interpretation of this tweet depends on the context in which Virgin was mentioned i.e. the regulation of access to specific sites.

`@Liberationtech: Pirate Bay blockade begins w/ Virgin Media http://t.co/fDtAkRFx by @JoshHalliday // it started with @eircom 3yrs ago!`

Figure 3.7: An information sharing tweet to the Eircom account.

3.5.3.2 Opinions/Complaints

Where Named Entities and Bag-Of-Words are sufficient

Consider the tweets in Figures 3.8 and 3.9, to the Orange UK and AT&T accounts, respectively. Each compares the organisation who owns the account to a competitor; favourably in the case of 3.8 and unfavourably in the case of 3.9.

```
@OrangeHelpers Vodafone scrapped £99 4 iphone 4S 16gb £36
tariff! Shame no price match, Vodafone signal terrible,
Orange is 3G in my area.
```

Figure 3.8: An opinion to the Orange UK account.

```
AT&T sucks! Been customer for 4 yrs & they just took
my unlimited data plan. Anyone have exp. w/ Verizon or
sprint? cc: @ATTCustomerCare
```

Figure 3.9: An opinion to the AT&T account.

Where Named Entities and Bag-Of-Words are insufficient

Consider the tweet in Figure 3.10, to the Comcast account. In this tweet, a user accuses NBC (the parent company of Comcast) of discriminatory practices. Note that the Named Entity and Bag-Of-Words representations identify NBC, but fail to represent the broader context in which NBC was mentioned.

```
Hey @comcastcares NBC’s hatemongering and race-baiting must
end! http://t.co/bCkL53n7 via @brentbozell please RT
```

Figure 3.10: An opinion to the Comcast account.

3.5.3.3 Statements and Random Thoughts

Where Named Entities and Bag-Of-Words are sufficient

Consider the tweets in Figures 3.11 and 3.12 to the Lexmark and Verizon accounts, respectively. 3.11 is a challenge to Lexmark by the user as to whether Lexmark will respond to the user’s comments. 3.12 praises Verizon, but argues that a competitor has more appealing products.

```
Guess I am going to find out if @LexmarkListens | Greetings
from Mexico
```

Figure 3.11: A statement to the Lexmark account.

```
@VZWSupport Yeah, I've been with you for 12 yrs.  I'm not
going anywhere anytime soon...But Boost Mobile has better
looking models.
```

Figure 3.12: A statement to the Verizon account.

Where Named Entities and Bag-Of-Words are insufficient

Consider the tweet in Figure 3.13, to the HP account. This tweet is a quote from Peter Drucker, a writer in the subject of management. The Bag-Of-Words representation identifies 'Peter' and 'Drucker'. The Named Entity representation identifies 'Peter Drucker', but fails to capture that the tweet is not about Drucker per se; rather it references a concept articulated by him.

```
RT @HPSupport:  "Management is doing things right,
leadership is doing the right things" - Peter Drucker #HPS3
```

Figure 3.13: A statement to the HP account.

3.5.3.4 Me Now

Where Named Entities and Bag-Of-Words are sufficient

Consider the tweets in Figures 3.14 and 3.15 to the Microsoft and Verizon accounts, respectively. Each expresses a general user sentiment about the company, with 3.14 being positive and 3.15 being negative.

@MicrosoftHelps @fj2k3films love windows 8 amazing keep finding new things stable fast can tot wait foe new EI hope adobe helps with flas

Figure 3.14: A ‘Me Now’ tweet to the Microsoft account.

@motodev @motorola @Verizon @VZWSupport Motorola and Verizon, I’m done. Getting non-Moto phone and leaving Verizon. Latest update worthless.

Figure 3.15: A ‘Me Now’ tweet to the Verizon account.

Where Named Entities and Bag-Of-Words are insufficient

Consider the tweet in Figure 3.16 to the Eircom account. In this tweet, a user complains that the poor quality of Eircom’s service has affected their ability to watch a movie. The Named Entity representation correctly identifies the movie (Alvin and the Chipmunks), whereas the Bag-Of-Words representation captures only the word ‘Alvin’. However, the Named Entity representation fails to capture the fact the subject of the tweet is in fact the Eircom network.

Dammit @eircom with your **** ADSL2 rollout and @VodafoneIreland for not pushing it. You’ve just ruined my children’s Alvin enjoyment.

Figure 3.16: A ‘Me Now’ tweet to the Eircom account.

3.5.3.5 Question to Followers

Where Named Entities and Bag-Of-Words are sufficient

Consider the tweets in Figures 3.17 and 3.18 to the AT&T and Comcast accounts, respectively. Both tweets criticise the relevant company by putting a question to a broader audience.

Anyone else with @att service running abysmally slow
in #austin the last few weeks? Am I being throttled?
@ATTCustomerCare

Figure 3.17: A 'Question to followers' tweet to the AT&T account.

about to get rid of my @comcastcares service, any
recommendations for alexandria? how's verizon DSL?
#comcast

Figure 3.18: A 'Question to followers' tweet to the Comcast account.

Where Named Entities and Bag-Of-Words are insufficient

Consider the tweet in Figure 3.19 to the Verizon account. In this tweet, the user encourages other users to 'twitter bomb' the Verizon account. While the Named Entity and Bag-Of-Words representations correctly identify 'twitter' and 'verizon' in this tweet, they fail to represent a key subject of the tweet i.e. 'Twitter bombing'[6].

can we twitter bomb @verizonwireless and @VZWsupport about
this? http://t.co/K3sEbl0y

Figure 3.19: A 'Question to followers' tweet to the Verizon account.

3.5.3.6 Presence Maintenance

Where Named Entities and Bag-Of-Words are sufficient

Consider the tweets in Figures 3.20 and 3.21 from the Sony and Dell accounts, respectively. Each of these tweets advertises the company's presence on a particular social media platform.

[6] https://en.wikipedia.org/wiki/Twitter_bomb

Visit the SonyListens YouTube channel at http://t.co/1mXZOgIK for great tips, tricks and troubleshooting videos on your Sony Electronics

Figure 3.20: A ‘Presence Maintenance’ tweet from the Sony account.

The Dell Channel Daily is out! http://t.co/CKlvO8FN ? Top stories today via @DellCares @lindaatdell

Figure 3.21: A ‘Presence Maintenance’ tweet from the Dell account.

Where Named Entities and Bag-Of-Words are insufficient

Consider the tweet in Figure 3.22 from the HP account. In this tweet the company promotes the different social media services users can use to interact with the company. This is done by rewording a well known expression. While the Named Entity and Bag-Of-Words representations identify the location (‘Vegas’) and the different social media services, they do not identify the true subject of the tweet.

@HPSupport: "What happens in Vegas stays on YouTube, Flickr, Twitter, Facebook and Google+" #goodtoknow #HPS3

Figure 3.22: A ‘Presence Maintenance’ tweet from the HP account.

3.5.3.7 About addressee

Where Named Entities and Bag-Of-Words are sufficient

Consider the tweets in Figures 3.23 and 3.24 to the Comcast and Dell accounts, respectively. Each tweet asks about using the organisation’s products in conjunction with the products of another organisation.

@comcastcares im noticing that i can get HBO on demand, but HBO West doesnt show up live.

Figure 3.23: An ‘About Addressee’ tweet to the Comcast account.

@DellCares I have an Inspiron 1420 with a seemingly faulty Nvidia GPU. How can I get this resolved?

Figure 3.24: An ‘About Addressee’ tweet to the Dell account.

Where Named Entities and Bag-Of-Words are insufficient

Consider the tweet in Figure 3.25 to the Microsoft account. In this tweet the user explains that they have tried resolving a problem with a Microsoft team in Asia but with no success. The user then asks to speak with a USA-based team. While the Named Entity and Bag-Of-Words representations identify 'Asia' and 'USA' they fail to identify the underlying subject linking these entities. That is, the user who authored the tweet is claiming that some characteristic of the Asian team is lacking, and that this prevents the user's query from being resolved.

```
@MicrosoftHelps Support Desk in Asia has been unable to
provide any solutions to our enterprise problem.  Possible
to escalate to USA?
```

Figure 3.25: An 'About Addressee' tweet to the Microsoft account.

3.5.3.8 Anecdote (me)

Where Named Entities and Bag-Of-Words are sufficient

Consider the tweets in Figures 3.26 and 3.27 to the Dell and Microsoft accounts, respectively. 3.26 describes a story of how a user fixed a problem themselves and how Dell then requested payment. In 3.27 the user describes how their laptop was stolen; and with it, the user's product activation code.

```
@DellCares I had my laptop completely re-done by my
brother.  Dell tried to charge him $150 for direction to
re-set it.  Ridiculous.
```

Figure 3.26: An 'Anecdote (me)' tweet to the Dell account.

```
@MicrosoftHelps my laptop was stolen.  I'd purchased
Microsoft office 2010.  I no longer have microsoft product
code.
```

Figure 3.27: An 'Anecdote (me)' tweet to the Microsoft account.

Where Named Entities and Bag-Of-Words are insufficient

Consider the tweet in Figure 3.28 to the Yahoo account. In this tweet, the user claims that their Twitter account was suspended because they had complained about racism on Yahoo fora. While the Named Entity and Bag-Of-Words representations identify Twitter and Yahoo, they do not identify the broader subject of the tweet i.e. the accusations of racism and censorship.

@YahooCare My twitter account @RarelySpoken suspended because I complained of the vile, ignorant and illegal racism on Yahoo!

Figure 3.28: An ‘Anecdote (me)’ tweet to the Yahoo account.

3.5.3.9 Anecdote (others)

Where Named Entities and Bag-Of-Words are sufficient

Consider the tweets in Figures 3.29 and 3.30 to the Orange UK and LG accounts, respectively. In 3.29, a user provides a story about their friend’s issue with trying to get a phone unlocked. In 3.30, a user tells a story about their friend needing to have their phone repaired.

@OrangeHelpers Hello my mate has got an iPhone 4 on orange how could she get it unlocked.Got the phone ebay.

Figure 3.29: An ‘Anecdote (others)’ tweet to the Orange UK account.

@LG_Service Hi, My friend’s Optimus 3D needs to be repaired. Here in Saudi Arabia there is no service center, where to repair it?

Figure 3.30: An ‘Anecdote (others)’ tweet to the LG account.

Where Named Entities and Bag-Of-Words are insufficient

Consider the tweet in Figure 3.31 to the Yahoo account. In this tweet, a user relays a racist posting while making the point that this is a continuation of a trend. Note that while the Named Entity and Bag-Of-Words representations identify ‘Yahoo’ and ‘London’, they fail to identify the broader context of the tweet, which is the user’s assertion that Yahoo has a persistent problem with racist posting from users.

@YahooCare Another Yahoo racist Brian • London, England • 1 hour 17 minutes ago Report Abuse ANOTHER F--G INDIAN

Figure 3.31: An ‘Anecdote (others)’ tweet to the Yahoo account.

3.5.3.10 Self promotion

Where Named Entities and Bag-Of-Words are sufficient

Consider the tweets in Figures 3.32 and 3.33 from the O2 and Comcast accounts, respectively. In 3.32, O2 promote themselves through their support of the Irish national rugby team. In 3.33, Comcast promote themselves through a statement about upcoming services.

```
RT @O2IRL: Ireland play England March 17th & you could have
your name on the back of a player's jersey with O2 Treats.
Enter here http://t.co/VaOdomJi
```

Figure 3.32: A ‘Self promotion’ tweet from the O2 account.

```
HBO GO on Xbox 360 Coming Soon for Xfinity Customers
http://t.co/8AQU85cB
```

Figure 3.33: A ‘Self promotion’ tweet from the Comcast account.

Where Named Entities and Bag-Of-Words are insufficient

Consider the tweet in Figure 3.34 from the Verizon account. In this tweet, Verizon promote their service by re-purposing a quote from Mark Twain. The Bag-Of-Words representation captures ‘Mark’ and ‘Twain’. The Named Entity representation identifies ‘Mark Twain’. However, the Named Entity representation does not capture the fact that the true interpretation is not Mark Twain himself, but the use of one of his quotes.

```
Kindness is the language that the deaf can hear & the blind
can see - Mark Twain.  Here to provide the kindest customer
service until 10PM.
```

Figure 3.34: A ‘Self promotion’ tweet from the Verizon account.

3.5.3.11 Exhort

Consider the tweets in Figures 3.35 from the O2 account and 3.36 to the AT&T account, respectively. In 3.35, O2 encourage users to follow their Twitter account by offering a chance to win a phone. In 3.36, a user encourages AT&T to improve their Twitter customer support behaviour.

```
@O2IRL: Samsung Night Run this Sun.  RT & follow to be
in to win a Samsung Galaxy Note.  Get 20% off selected
Samsung, http://t.co/4fXhGXhV
```

Figure 3.35: An ‘Exhort’ tweet from the O2 account.

```
Hey @att @attnews @attcustomercare you could pick up a
few tips on Twitter responsiveness from @DeltaAssist
@verizonsupport @verizonwireless
```

Figure 3.36: An ‘Exhort’ tweet to the AT&T account.

Where Named Entities and Bag-Of-Words are insufficient

Consider the tweet in Figure 3.37 to the Vodafone account. In this tweet a user asks Vodafone to sever their ties with Rupert Murdoch. This tweet references a scandal at the time involving the (now defunct) News of The World newspaper, which was owned by Murdoch. The Bag-Of-Words representation identifies ‘Rupert’ and ‘Murdoch’. The Named Entity representation identifies ‘Rupert Murdoch’, but fails to capture the context in which he was mentioned i.e. the hacking scandal involving his newspaper.

```
@VodafoneUK Please don’t encourage Rupert Murdoch for
hacking and bribery by buying ads in his new newspaper
#rupertmurdoch #Vodafone
```

Figure 3.37: An ‘Exhort’ tweet to the Vodafone account.

3.5.3.12 Media Use

Where Named Entities and Bag-Of-Words are sufficient

Consider the tweets in Figures 3.38 and 3.39 to the Nokia and Microsoft accounts, respectively. In 3.38 a user complains about the functionality of Youtube on their phone. In 3.39 a user praises Windows 8 in general while lamenting the functionality of Netflix.

```
@NokiaHelps and what should I do about my YouTube..it’s not
opening after an update from store I reinstall this but d
prblm is same
```

Figure 3.38: A ‘Media Use’ tweet to the Nokia account.

```
@BuildWindows8 @MicrosoftHelps been using #windows8 for
about a week now and i love it.  if Netflix worked it would
be 10 out of 10
```

Figure 3.39: A ‘Media Use’ tweet to the Microsoft account.

Where Named Entities and Bag-Of-Words are insufficient

Consider the tweet in Figure 3.40 to the Comcast account. In this tweet the user describes difficulty they had obtaining HBO service by making reference to a now-renamed Comcast policy called ‘Make it right’[7]. Although the Named Entity and Bag-Of-Words representations capture ‘HBO’ they fail to capture the reference to the customer support policy.

```
@comcastcares Hey Bill, is 3 months of HBO for 3 long bus
trips to the Comcast store as a result of employee apathy
Making It Right"" ?
```

Figure 3.40: A ‘Media Use’ tweet to the Comcast account.

3.6 Discussion

The various examples presented in the previous section reveal a key limitation of the Named Entity and Bag-Of-Words representations. The inability of each of these representations to capture contextual information limits their ability to provide a comprehensive understanding of different forms of user expression. Note that this limitation is not restricted to a entity type, or a single form of context. In the case of organisation entities, unidentified contexts include accusations of unethical behaviour (as in Figure 3.31) and a core part of the company’s service (as in Figure 3.16). In the case of location entities, unidentified contexts include turns of phrase (as in Figure 3.22) and different levels of ability between teams in different regions (as in Figure 3.25). In the case of person entities, unidentified contexts include ongoing events in which the person is involved (as in Figure 3.37) and the fact that a person’s quote was the

[7] https://www.theverge.com/2014/8/7/5971857/we-re-on-it-comcast-customer-service-employees-cards

subject of the tweet rather than the person themselves (as in Figure 3.34).

These results indicate that the lack of context in representations such as Named Entities and Bag-Of-Words does not prevent them from representing certain forms of user expression. The examples in the previous section show that the Named Entity and Bag-Of-Words were able to represent some forms of each of the different types of user expression. However, these representations were not able to *fully* represent these forms of user expression. Therefore, with respect to the research question of this experiment, these results indicate that Named Entities and Bag-Of-Words are insufficient for completely representing different forms of user expresion.

3.7 Conclusion

This chapter described an experiment investigating how the absence of context in representations such as Named Entities and Bag-Of-Words limits their ability to represent different forms of user expression. The results of this experiment indicate that the fact that these representations do not capture context makes them insufficient for fully reprsenting these different forms of user expression. Recall from Chapter 1 that a common approach for providing the context that is lacking in representations such as Named Entities and Bag-Of-Words is to leverage the conceptual hierarchy in an external linguistic resource. Recall also that there are different types of linguistic resource, raising the question of how to choose a resource. Chapter 1 described two criteria that can aid this choice: social relevance and the extent to which the resource reflects domain category distinctions. The investigation of the former criterion is the subject of the next chapter. The latter criterion will be investigated in Chapter 5.

Chapter 4

A Comparison of External Linguistic Resources

The experiment described in the previous chapter revealed a limitation of the Named Entity and Bag-Of-Words representations: the fact that these representations do not capture contextual information prevents them from fully representing different forms of user expression. This limitation can be overcome by leveraging the conceptual information in an external linguistic resource. Chiarcos et al. divide such resources into two types, distinguishing between "lexical resources in a strict sense (which provide specifically linguistic information, e.g., grammatical features, as found, e.g., in a dictionary, or in a WordNet), and general knowledge bases (such as classical thesauri or semantic repositories such as YAGO and DBpedia)"[23]. Each of these types of resource provides a different kind of contextual information. The context provided by lexical resources pertains specifically to grammatical relations, whereas the context provided by general knowledge bases pertains to principles of knowledge organisation. Recall the example provided in Chapter 1 of the difference between the way 'law' is hierarchised in WordNet and DBpedia. WordNet defines the parent of 'law' as 'collection', stressing the grouping property of the term. DBpedia defines the parent of 'law' as 'justice', stressing the legal and judicial property of the term. These different forms of context will naturally yield structured representations that capture different kinds of information. It is therefore necessary to be able to compare different linguistic resources in order to determine which is most suitable for the personalisation task at hand. Recall also from Chapter 1 that Bentivogli et al. argue that the criterion of social relevance facilitates the evaluation of the suitability of a linguistic resource.

This chapter describes an experiment that compares a representative example of a purely lexical resource with a representative example of a general knowledge base. WordNet is chosen as an example of a purely lexical resource. The author argues that WordNet is a representative example of a purely lexical resource as it is widely used in NLP research [81], [82]. DBpedia is chosen as an example of a general knowledge base. The author argues that DBpedia is a representative example of a general knowledge base because, as stated in Section 3.4, the quality of its knowledge rivals that of Encyclopedia Britannica [79]. These resources will be compared on the basis of their social relevance. This comparison will investigate whether the social relevance of a resource influences its ability to represent user information. For descriptions of WordNet and DBpedia, please refer to Section 2.3. To ensure this comparison is format-independent (e.g. does not focus solely on the tweet format) this study investigates Twitter and LinkedIn.

Section 4.1 describes the objective of this experiement. **Section 4.2** describes the case that will be investigated in this experiment. **Section 4.3** describes the research question of this experiment. **Section 4.4** describes the process for generating the structured representations used in this experiment. **Section 4.5** describes the experiment design. **Section 4.6** discusses the results of the experiment and **Section 4.7** concludes.

4.1 Objective of this experiment

Recall that the second objective of the research question of this thesis is the comparison of different kinds of external linguistic resources in providing the context that is absent from content-based representations such as Named Entities and Bag Of Words. Recall from Section 2.6.5 that the experiment described in this chapter must assess the ability of each resource to represent user interests and knowledge. Recall also from Section 2.6.5 that the social relevance of each resource must be considered when making the comparison between linguistic resources. The objective of the experiment described in this chapter is thus to investigate the relationship between each resource's social relevance and its efficacy in: (i) Comparing users' self-descriptions of their interests and knowledge on different social media profiles (ii) Determining whether information from one profile can be used to augment information for another. This experiment will examine how the social relevance of a resource influences its ability to perform each of these tasks.

4.2 Case Description

Each of the social media services Twitter and LinkedIn services allows users to describe their interests and knowledge by: (i) Filling in profile information (ii) Posting status updates. However, the percentage of users who post status updates on LinkedIn is significantly lower than the percentage of users who do so on Twitter [83]. On the other hand, LinkedIn users fill in far more of their profiles on average than Twitter users [28].

Given the different ways in which users use these services, it is possible that they provide different representations of their interests and knowledge on each one. For example, a user may indicate a new-found interest in Linguistics through their tweets before they list this subject on their LinkedIn profile. The experiment described in this chapter examines the relationship between users' descriptions of their interests and knowledge on each service.

The author argues that this case meets the requirements of Case Study 2 described in Section 2.6.5 for the following reasons:

1. It involves generating separate representations of user interests and knowledge for each of the Twitter and LinkedIn services.
2. It involves combining and contrasting these separate representations.

This experiment will investigate the efficacy of each of WordNet and DBpedia in performing both of the above tasks. Thus, this experiment meets the requirements described in Section 2.6.5.

4.3 Research Question of this Experiment

The research question of this experiment is as follows:

How does the social relevance of a linguistic resource influence its ability to represent user interests and knowledge?

This question is divided into two parts:

1. Comparing the extent to which a user's description of their interests and knowledge on Twitter corresponds with their description of these characteristics on LinkedIn.
2. Determining whether information obtained from a user's tweets can be used to recommend terms for that user's LinkedIn profile.

This will allows for a full examination of the relationship between the social relevance of a linguistic resource and that resource's ability to describe user interests and knowledge, thereby achieving the objectives described in Section 4.1.

4.4 Generating the structured representation

Texts (i.e. tweets, LinkedIn descriptions) are processed, and the WordNet and DBpedia resources accessed, using the Pattern NLP library for Python [52]. This library is described in Section 2.5.

Ma et al. discuss methods for leveraging hierarchical links in external resources to represent user interests [76]. The authors make a distinction between implicit and explicit interests. Explicit interests are identified by linking texts a user has written with a label in a hierarchy. Implicit interests are then identified by obtaining the subtype and/or supertype of the identified label. For example, consider a hierarchy in which 'Knowledge Representation' is the supertype of 'Semantic Web'. If a user explicitly indicates they are interested in 'Semantic Web', they are implicitly indicating that they are interested in 'Knowledge Representation'. However, Ma et al. provide the caveat that only the immediate supertypes and subtypes (i.e. one level above/below) should be identified for a particular label. In the study described in this chapter, implicit information is obtained by identifying supertype labels only. This decision was taken with reference to the 'is-a' subsumption relation between types[1]. Suppose the 'is-a' relation holds between two subtypes B, C and their supertype, A. Any instance of B is of type A, as is any instance of type C. However, the reverse does not hold i.e. an instance of type A is not necessarily an instance of type B, as it could be an instance of type C. Indeed, A may not be an instance of either B or C. For example, if a user explicitly expresses an interest in 'Knowledge Representation' they are not necessarily implicitly expressing an interest in its subtype, 'Semantic Web'.

In order to apply Ma et al.'s approach, two questions must be answered:

1. How will 'label' be defined with respect to WordNet/DBpedia?

2. How will the hierarchical relations between labels be defined with respect to WordNet/DBpedia?

Each of these questions is addressed in the following paragraphs.

The Extended WordNet Domain labels created by González-Agirre et al. [84] will be used as labels with respect to WordNet. These are subject labels assigned to WordNet synsets e.g. 'sports', 'medicine' etc. In WordNet, the hypernymy/hyponymy relation links more general

[1] The is-a relation is described here: `http://geneontology.org/page/ontology-relations#isa`

synsets with more specific synsets [85]. For example, ‘furniture’ is a hypernym of ‘bed’, and ‘bed’ is a hyponym of ‘furniture’. Thus the hypernym/hyponym relation will be used as hierarchical relations with respect to WordNet.

As described in Section 2.3, DBpedia categories represent collections of similar subjects. Thus, DBpedia categories will be used as labels with respect to DBpedia. The Simple Knowledge Organisation System (SKOS)[2] ‘broader’ relation is used as a hierarchical relation with respect to DBpedia. As described in Section 2.3, the ‘broader’ relation links a subject with an object that is semantically more general. Its inverse is the ‘narrower’ relation. For example, ‘broader’ holds between ‘Semantic Web’ and ‘Data Management’ and ‘narrower’ holds between ‘Data Management’ and ‘Semantic Web’. The decision to use the ‘broader’ relation is made with reference to the approach adopted by Forsberg and Borin in creating their Swedish FrameNet++ lexical framework[86]. Forsberg and Borin use the ‘broader’ and ‘narrower’ relations to express hypernymy and hyponymy, respectively.

Structured representations of a user’s tweets and LinkedIn information are generated by the following three steps:

1. Keyphrase identification. A keyphrase is a single- or multi-word phrase that “expresses the primary topics and themes” of a text [87].

2. Linking keyphrases to Extended WordNet Domain labels/DB category labels to form a term-list representing the user’s interests and knowledge. Central to this step is the task of Word Sense Disambiguation (WSD), which automatically identifies the correct meaning for a word in text. A description of WSD, and the WSD methods employed in this case study with respect to WordNet and DBpedia, can be found in Appendix B.

3. Calculating term weights.

These steps are described in the following subsections.

4.4.1 Keyphrase Identification

The user’s LinkedIn lists of skills, interests and courses are treated as lists of keyphrases with respect to each of DBpedia and WordNet. However, the process for identifying keyphrases in text (e.g. a textual description on LinkedIn, a tweet) differs for each resource. For the DBpedia

[2] http://www.w3.org/2004/02/skos/

category-based representation candidate keyphrases derive from a precompiled list i.e. the controlled vocabulary created using Mihalcea and Csomai's Word Sense Disambiguation method, described in Section B.2. In the case of the Extended WordNet Domains-based representation, no such list exists, meaning a different approach must be used for identifying keyphrases. Riloff describes how keyphrases can be identified by the use of *case frames* [88]. A case frame is a text pattern consisting of a verb and one or more empty 'slots' that can be searched in text in order to find noun phrases that are important for the text i.e. keyphrases[3]. For example, the frame 'X murdered Y' has two empty slots, X the perpetrator and Y the victim. The grammatical roles of X and Y are subject and object, respectively. Using this case frame, a text can be searched to identify the perpetrators (subjects) and victims (objects) of murder. A case frame approach cannot be adopted in the study described in this chapter as it relies on previous domain knowledge of the text being processed. For example, in order to construct the case frame 'X murdered Y', one would have to know that the text being analysed contained references to murder. Instead, each text is analysed in order to extract its noun phrases, whose role is either subject or object. These noun phrases are then sub-divided into contiguous sequences of words of length n, called *n-grams*. In the study described in this chapter n is a value between one and the length of the noun phrase, inclusive. For example, if the noun phrase 'Natural Language Processing' has been identified, the associated n-grams are 'Natural', 'Language', 'Processing', 'Natural Language', 'Language Processing' and 'Natural Language Processing'. This is similar to the approach of Chang et al. who decompose user queries into n-grams as part of a method for automatically inferring user query intent [89]. Note that this approach is implicitly adopted in the DBpedia category-based approach. For example, although 'Natural Language Processing' is a keyphrase in Wikipedia, so too are 'Natural Language', 'Natural', 'Language' and 'Processing'.

4.4.2 Linking keyphrases to labels

In order to link a keyphrase with an Extended WordNet Domains label/DBpedia category label, the keyphrase must first be linked with a WordNet synset/DBpedia resource URI. The processes by which this is achieved for LinkedIn skill, interest and course lists and for texts (i.e. tweets, LinkedIn textual descriptions) are given below.

[3] Recall that a noun phrase consists of a noun which may be modified by other words, for example 'A tall man'.

4.4.2.1 LinkedIn skill, interest and course lists

Magnini et al's WSD method, described in Section B.1, offers a means of linking keywords in a user's LinkedIn skill, interest and course lists to Extended WordNet Domain labels/DBpedia category labels. Unambiguous keyphrases in these lists can be linked directly with a WordNet synset or DBpedia URI. These unambiguous keyphrases can then provide a basis for interpreting ambiguous terms. For example, the unambiguous 'XML' can be linked with the 'Computer Science' domain, providing a basis for interpreting the ambiguous 'Java'.

4.4.2.2 Texts

Terms in text are disambiguated with respect to WordNet synsets and DBpedia resource URIs using the Word Sense Disambiguation algorithms described in Appendices B.1 and B.2, respectively. These particular WSD algorithms are chosen because each uses a multi-domain corpus as the basis for performing disambiguation. A fundamental aspect of Magnini et al.'s approach is that it factors the domain of a particular piece of text into the disambiguation of ambiguous phrases in that text. In the case of Mihalcea and Csomai's approach, the Wikipedia corpus is leveraged in disambiguating phrases. Since the text analysis in the study described in this chapter will address texts from a variety of domains, the varied nature of the corpora employed in both WSD approaches makes them ideal candidates.

Tweet preprocessing

Before tweets are analysed, they are preprocessed as follows:

- The word 'RT' appearing at the beginning of a tweet (indicating the tweet is a retweet) is removed.
- Characters repeated consecutively more than twice are replaced with two consecutive characters (e.g. 'gooood' becomes become 'good') as in Vosecky et al. [90].
- For hashtags the '#' symbol is removed and the tag is split using the capital letters, as in [91]. For example, '#MachineLearning' becomes 'Machine Learning'.
- Twitter usernames are replaced with a generic string 'AT_USER'[4]. If this is not performed problems could occur in interpreting subject or object noun phrases. Consider the following sample tweet 'What's up spartan?'. The chunks[5] produced by Pattern library for this tweet can be seen Table 4.1[6]. This chunking is problematic because the text 'spartan'

[4] In Bifet and Frank's study the string 'USER' is used.

[5] For a description of chunking please see Section 2.5.

[6] 'What' does not appear as it is a stop word.

can be linked to the WordNet synset ‘Spartan’, a resident of Sparta. This synset has the domain ‘history’, which is not necessarily relevant for the tweet. It could be the case that the user ‘@Spartan’ has an interest in Sparta and Spartans, but making this determination requires unavailable external knowledge. Replacing ‘spartan’ with ‘AT_USER’ addresses this issue, as there is no WordNet synset for the text ‘AT_USER’.

Chunk	**Chunk tag**	**Tag meaning**	**Role**
[‘’s’, ‘up’]	VP	Verb phrase	None
[‘@’,‘spartan’]	NP	Noun phrase	Object

Table 4.1: A chunking of the sample tweet ‘What’s up @spartan?’.

However, the tweet format could potentially still pose problems for each of the DBpedia and WordNet approaches. Misspellings, slang etc. could result in certain phrases not being disambiguated in the DBpedia approach. This could be, for example, because not enough words in the tweet were found in a Wikipedia article to allow for disambiguation. Misspellings are specifically accounted for in the WordNet approach, as the POS-tagger employed in the Pattern library has a robust strategy for handling un-recognised words [53]. Poor sentence construction (e.g. improper punctuation) could pose a problem for the WordNet approach as it could result in subject and/or object chunks not being identified. This would be because chunking naturally depends on sentence structure. This would not pose a problem for the DBpedia approach, as it does not consider sentence structure. The key point to note is that the evaluation method of this study, described in Section 4.5.2, does not penalise an approach for failing to identify a term in the user’s tweets or LinkedIn profile information. Each approach is evaluated on the basis of its ability to compare and combine information between the user’s tweets and LinkedIn information. Thus, the aforementioned issues with the tweet format will not affect the evaluation of each external linguistic resource.

Unambiguous keyphrases identified in text are linked directly. The process for linking ambiguous keyphrases to labels is described in the following subsections.

Extended WordNet Domains

Magnini et al.’s algorithm is employed to link keywords to their appropriate WordNet synset [92].

Once a WordNet synset has been identified from a keyphrase the domain label for this synset as well as the domain labels for its hypernyms are obtained. The following domains are ignored:

1. Factotum
2. Number
3. Color
4. Time Period
5. Person
6. Quality
7. Metrology
8. Psychological Features

These domains are overly generic. Each of the domains 2-8 is a sub-domain of the 'Factotum' domain, which is applied to synsets for which no specific domain can be found.

Separate corpora are used for processing tweets and for processing LinkedIn textual data. For the former, a 36.4 million-word tweet corpus is used. For the latter the approximately 300 million-word blog corpus compiled by Schler and Koppel [93] is used. This corpus is deemed suitable given both its size and the fact that it contains posts on a wide variety of topics, for example: 'Real Estate', 'Arts', 'Education', 'Engineering', 'Law' etc.

Magnini et al report that a context of at least 100 words should be used to disambiguate an ambiguous phrase (50 words before the word and 50 words after). For tweets, as such a context is not available, the whole tweet is used. In an analysis conducted on 1.5 million random tweets in 2009 the Oxford University Press' Dictionary team found the average number of words per tweet to be 14.98. [94].

DBpedia

Candidate keyphrases are ranked using the Keyphraseness measure described by Mihalcea and Csomai, described in Section 4.5.2. A number of top ranked keywords is chosen such that the ratio of keyphrases to words in the text is 6%. The WSD method of Mihalcea and Csomai is then used to link keywords to their appropriate Wikipedia article, W[7]. Wikipedia article

[7] Described in Appendix B.2.

titles are used to derive DBpedia resource URIs. Once a DBpedia URI has been identified from a keyword this URI is then linked to the DBpedia category bearing the same label using the Dublin Core 'subject'[8] relation. If no such category exists (for example in the case of the programming language Prolog), this means that W did not meet the criteria required to be given its own category[9]. In this case, all categories related to the DBpedia resource URI by the 'subject' relation are obtained. Once a DBpedia category URI has been identified its label is obtained, as are the labels of the categories related to the URI by the SKOS 'broader' relation.

Keyphrases relating to categories of the following types are ignored:

- Person
- Place
- Organization
- Event
- Animal
- Film
- Television Show
- Book
- Play (Theatre)

This is because in a list of top LinkedIn skills compiled by LinkedIn Profile Services[10], not a single entity of these types appears. The DBpedia category 'Main topic classifications' is not considered as it is a table of contents for other categories. Similarly, DBpedia categories such as 'Wikipedia categories named after information technology companies of the United States' are not considered as these refer specifically to the way in which the Wikipedia hierarchy is organised, rather than the categories in it.

Mihalcea and Csomai define the context of a keyphrase to be the paragraph in which it appears. For the DBpedia approach, this context is defined in the same way as in Magnini et

[8] `http://dublincore.org/documents/2012/06/14/dcmi-terms/?v=terms\#subject`

[9] Guidelines for creating Wikipedia categories can be found here: `http://en.wikipedia.org/wiki/Wikipedia:Categorization`

[10] Available at: `http://linkedinprofileservice.co/linkedin-profile-tips-advice/linkedin-skills-list/`

al.'s study i.e. 50 words before and 50 words after the ambiguous keyphrase. The reason for this is that Magnini et al. determined this context through experiments on Semcor[11]. Semcor is derived from the Brown corpus, which is a balanced corpus of English written texts. The author determines a context length identified with reference to this corpus to be appropriate in analysing users' various LinkedIn texts.

Any links appearing in text are extracted. The contents of the linked page are obtained and cleaned of HTML using the AlchemyAPI text extraction tool[12]. The remaining text is analysed in the same manner as the other texts analysed in this study (e.g. tweets, textual descriptions on LinkedIn). However, an upper word limit must be placed on link content in order to avoid spending an inordinate amount of time analysing individual links. Kwak et al. conduct an investigation of the way information spreads on Twitter and find that 85% of topics discussed on Twitter relate to news events [95]. Thus, in the study described in this chapter, the upper word limit for links is determined with reference to ideal news article length. In a Reuters blog post on this subject, MacMillan writes: "Reuters editors see stories that exceed 500 or 600 words as indistinguishable from 'Gravity's Rainbow' [13] " [96]. Thus, the upper word limit for links is set at 550 words.

Recall from Table B.6 in Appendix B.2 that the Feature-Based disambiguation method achieves the highest F-measure. However, implementing this method for the experiment described in this chapter proved to be unworkable. Consider for example the term 'Xbox', which appears as a keyphrase in 4008 articles. To apply the Naïve Bayes approach, the feature set Mihalcea and Csomai describe would have to be gathered for every occurrence. Performing this task over tweets as well as LinkedIn profile information for multiple experiment participants proved to be prohibitively expensive in terms of time taken. The next highest WSD method with regard to F-measure is Most Frequent Sense (MFS). However, the study described in this chapter involves the analysis of content pertaining to users' areas of profession. This information will span a variety of subject areas, each of which can have subject-specific interpretations of terms. The adverse effect on performance of MFS that this can cause is shown by Ponzetto et al. [97]. In an evaluation of Word Sense Disambiguation approaches on Sports- and Finance-specific corpora, Most Frequent Sense achieves F-measures of 19.6 and 37.1, respectively. Thus, the Knowledge-based method is used to disambiguate keyphrases in the DBpedia approach. Its F-measure of 75.99 is

[11] `https://www.gabormelli.com/RKB/SemCor_Corpus`

[12] http://www.alchemyapi.com/api/text/urls.html

[13] A 760-page novel written by Thomas Pynchon.

higher than the 42.0 and 47.8 F-measures that Ponzetto et al.'s method achieves on the Sports- and Finance-specific corpora, respectively. This is also higher than the F-measure obtained by McCarthy et al. [98]. As well as this, the accuracy of the Knowledge-based approach is higher than that of approaches such as Agirre and Soroa [99] and Gomes et al. [100] Furthermore, Michelson and Macskassy employ the Knowledge-based method as part of their system, Twopics, that identifies users' topics of interest based on their tweets [21][14].

Another finding reported by Mihalcea and Csomai's evaluation is that 10% of keyphrases identified in test pages could not be disambiguated because they did not appear anywhere else in Wikipedia article text as links. This is most likely because links to the Wikipedia article appear using surface forms only e.g. the article 'Java (programming language)' may be linked using forms such as 'java', 'java programming language' etc. but not the exact text of the article title. If a candidate keyphrase does not appear in links in the text of a Wikipedia article, it does not have any possible sense definitions. This means it cannot be disambiguated. Thus, in the study described in this chapter, the controlled vocabulary used consists solely of surface forms found in the text of Wikipedia articles.

The fact that different methods of analysis are employed in linking keyphrases to DBpedia category labels and Extended WordNet Domain labels may lead one to believe that the resulting representations cannot be compared. However, similarly to Abel et al.'s study [7], the difference between the methods of analysis is a result of the differences between representations. The Entity-based, Topic-based and Hashtag-based models generated in Abel et al.'s study each represent different aspects of unstructured content. The methods required for generating each model are therefore necessarily different. The representations are compared by employing each in the same task. Similarly, as WordNet and DBpedia represent information differently, the methods of linking keyphrases with labels in each resource are different. However, the resulting structured representations will be compared in the same task.

4.4.3 Calculating term weights

The tf-idf weighting scheme is not used[15]. This is because tf-idf normalises a term's occurrence frequency in an individual document with its occurrence frequency in an entire collection. This means that terms that appear in a large number of texts would receive low weights, when the

[14] As noted in the introduction to Chapter 1, the reason Michelson and Macskassy's approach is not applied in the research described in this dissertation is it focuses solely on Named Entities in tweets.

[15] For a description of tf-idf please see Section 2.4

opposite should be the case. Consider for example if a user tweeted 100 times and the term ‘Linguistics’ appeared in each tweet. The inverse document frequency of ‘Linguistics’ would be 0. Multiplying term frequency by the above *idf* would give a weight of 0. Instead of tf-idf, a term’s weight is calculated using the frequency based method described in Section 2.4. For example, if there are a total of 4 term mentions and ‘Linguistics’ has been mentioned twice its weight will be 0.5. Terms with only a single mention are discarded before weights are calculated. This decision was taken in order to minimise noise in the form of outlying terms.

4.5 Experiment design

Eight users participated in the evaluation. Participants were identified by two means: (i) An email circulated in the research group in which the author works. (ii) Tweeting at Twitter users from the university in which the author works. The criteria for selection of participants were as follows:

- LinkedIn - Users who had logged in to their profile within a month of the date the email requesting participation was sent. This decision was made with reference to the LinkedIn usage statistics reported by Jeff Bullas [101]. Bullas reports that the average amount of time spent by users on LinkedIn in 2014 was 17 minutes per month. Thus, the author concludes that setting the time period any smaller than a month (e.g. a week) would be unnecessarily restrictive.
- Twitter - Users who have tweeted 1000 times or more. This requirement was determined with respect to the study described in Al Zamal et al. [102]. These authors aim to infer a user’s attributes from their tweets, and analyse 1000 tweets per user to do this.

While eight users is certainly a small number, the author makes the below argument to support the relevance of the results that derive from this experiment. Recall the following quote from Flyvbjerg, provided on page :

``from both an understanding-oriented and an action-oriented perspective,
it is often more important to clarify the deeper causes behind a given problem
and its consequences than to describe the symptoms of the problem and how
frequently they occur. Random samples emphasizing representativeness will
seldom be able to produce this kind of insight; it is more appropriate to
select some few cases chosen for their validity.''[32]

Recall also from Section 4.1 that the objective of the experiment described in this chapter is to fully explore the categorisation hierarchies of two different linguistic resources. Thus, the value of this experiment's participants lies in the nature of these participants' data and how well it facilitates this exploration. With this in mind, consider the fact that the average number of tweets for male and female Twitter users is 567 and 610, respectively [103]. This makes a user who has tweeted 1000 times or more quite unusual. Each participant in this experiment is therefore a case of a hyperactive Twitter user whose Twitter profile provides a wealth of information to be categorised by each resource. It is this wealth of information that makes the participants of this experiment significant.

4.5.1 Dataset Description

The user's 1000 most recent tweets are collected. As described in the previous section, this number was chosen with reference to the study conducted by Al Zamal et al.
The following information is collected from the user's LinkedIn profile:

1. The user's summary, as in Figure 4.1.
2. The user's skill, interest and course[16] lists, as in Figure 4.2 , 4.3 and 4.4 respectively.
3. The user's textual descriptions of their educational and professional experience, as in Figure 4.5 and 4.6, respectively.
4. The textual descriptions of LinkedIn groups to which the user belongs, as in Figure 4.7.

[16] Courses the user has studied.

```
I am currently working as a PhD student at the Knowledge
and Data Engineering Group (KDEG) at Trinity College
Dublin.  I also demonstrate for undergraduate courses
in Mathematics, Java, C++ and Prolog in Trinity College,
Dublin.  I will complete my studies in March 2016.
```

Figure 4.1: A paragraph from the author’s LinkedIn summary.

```
[‘Computer Science’, ‘Java’, ‘Python’]
```

Figure 4.2: A truncated version of the author’s LinkedIn skill list.

```
[‘Cinema’,‘Music’]
```

Figure 4.3: A truncated version of the author’s LinkedIn interest list.

```
[‘Prolog’, ‘Perl’, ‘C++’]
```

Figure 4.4: A truncated version of the author’s LinkedIn course list.

This degree involved the study of the ways in which images and text can be combined in order to convey meaning. Among the fields covered were cinematography, philately, poster art and illustration. My dissertation concerned the application of Peircean semiotic theory to musical phenomena.

Figure 4.5: A textual description of the author's M.Phil degree.

During this internship I worked on the user interface for the CloudWatch Console, which allows customers of Amazon Web Services to monitor the health of their services. My project during the internship was to integrate EC2 instance names in the console.

Figure 4.6: A paragraph from the LinkedIn description of an internship undertaken by the author.

This group is dedicated to the field of Computational Linguistics and other related fields, such as Natural Language Processing, Corpus Linguistics, etc.

Figure 4.7: A paragraph from the description for the 'Computational Linguistics' LinkedIn group.

4.5.2 Evaluation Method

The evaluation of this experiment consists of two parts:

- Assessing the social relevance of the DBpedia category hierarchy and the WordNet Domains hierarchy.

- Assessing the ability of each resource with respect to parts 1 and 2 of the research question of this experiment, given in Section 4.3.

4.5.2.1 Assessing the social relevance of each resource

To estimate the social relevance of WordNet Domains, Bentivogli et al. leverage the hierarchies of the Yahoo! and Google web directories. Since the Yahoo! and Google web directories are now defunct, the DMOZ[17] web directory is used in the experiment described in this chapter. The DMOZ hierarchy is a volunteer-edited project that categorises web links according to their subject matter. It is the largest such resource on the web, with approximately four million sites and over one million categories[18]. The comparison will examine two distinct categories in the DMOZ classification hierarchy that have corresponding WordNet Domains and DBpedia categories. For each of these categories, the DMOZ subcategories for which there are equivalent WordNet Domains and DBpedia categories will be examined. For each equivalent WordNet Domain, the parent domain will be identified. For each equivalent DBpedia category, the broader category will be identified (using the SKOS broader relation). This will allow for a comparison with the DMOZ hierarchy for each of WordNet Domains hierarchy and DBpedia category hierarchy. To avoid confusion between DMOZ categories and DBpedia categories, throughout this discussion DMOZ categories will be referred to as 'subjects'.

4.5.2.2 Comparing WordNet and DBpedia with respect to parts 1 and 2 of the research question

The experiment is split into two sessions. In the first session, the user logs in to their Twitter and LinkedIn accounts. Their data is then collected and structured representations are generated using the process described in Section 4.4. In the second session, the user is asked to provide subjective evaluations with regard to each part of the research question.

[17] https://www.dmoz.org/

[18] The DMOZ resource was made defunct on 17th March 2017. The static mirror of the DMOZ site lists 3,861,210 sites listed under 1,031,722 categories. The static mirror can be found here: http://dmoztools.net/

A term's weight in the LinkedIn or Twitter term-list is directly related to the total number of term mentions in that list. As this total can differ between LinkedIn and Twitter, comparisons between term-lists cannot be made using weights. Term ranks are used instead.

It must be noted that the number of participants in this experiment was quite small. As such, the results presented here should be considered indicative rather than definitive.

Procedure for part 1 of the research question

For Question 1 the relative ranks of terms that appear in both the Twitter and LinkedIn term-lists are compared.

The participant is shown a series of assertions about the relative ranks of terms that appear in both their Twitter and LinkedIn term-lists. For example, if 'Linguistics' is the third ranked term in the user's LinkedIn term-list but the fifth ranked term in their Twitter term-list, the user is shown a statement asserting that 'Linguistics' has less prominence in their tweets than in their LinkedIn profile. The participant can answer affirmatively or negatively to each assertion. However, if the analysis has incorrectly identified a term, the participant can indicate this instead of responding. They can also indicate that the term denotes a subject that was relevant for them in the past, but is not anymore.

If a term appears in the participant's LinkedIn term-list but not in their Twitter term-list, the participant is shown a statement asserting that the term has less prominence in their tweets than in their LinkedIn profile.

Procedure for part 2 of the research question

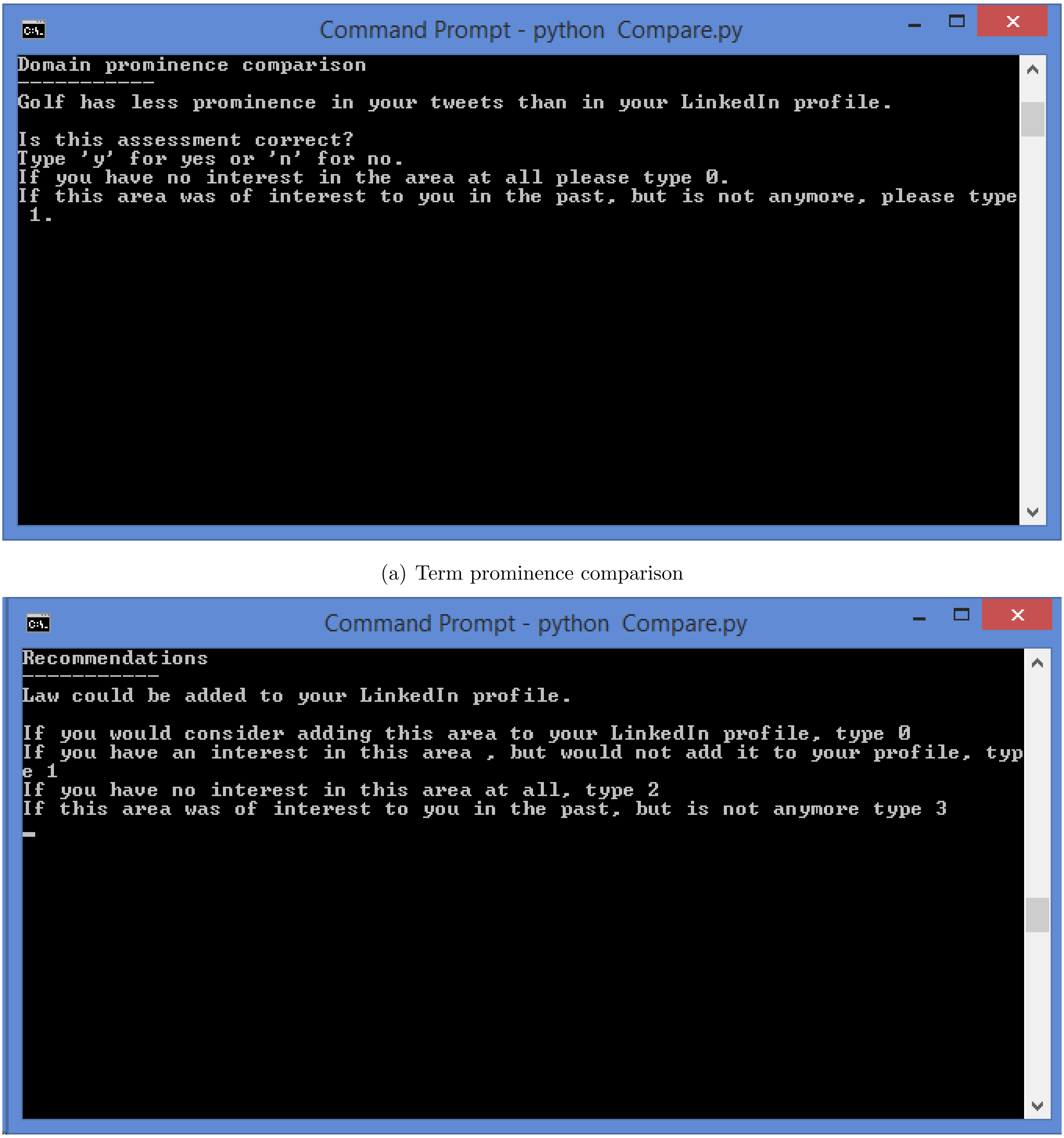
For Question 2, only Twitter terms whose rank is equal to or higher than the lowest ranked term in the LinkedIn term list are recommended. For example, if the participant's LinkedIn term-list contains six ranks, only Twitter terms of rank six or higher are recommended. If no terms are found in the participant's LinkedIn profile, only the first-ranked Twitter term(s) is recommended.

The participant is shown a series of recommendations for their LinkedIn profile. The participant can answer affirmatively or negatively to each recommendation. Alternatively, they can indicate that although the term denotes an area that is relevant for them they would not add it to their LinkedIn profile. This could be, for example, because they do not want to list the term on their professional profile or they do not feel sufficiently confident in their knowledge of the subject the term denotes. They can also indicate that the term denotes a subject that was relevant for them, but is not anymore.

Allowing participants to indicate terms were relevant for them in the past ensures that the efficacy of the Extended WordNet Domain-based and DBpedia category-based representations is assessed solely with regard to the terms the representations contain. For example, suppose a representation contains a term correctly inferred from a tweet written a year ago. The fact that the term is not relevant for the participant anymore should not be counted against the representation.

For the DBpedia category-based representation, terms in the format 'Branches of X' are presented to participants as 'X', as these pages contain lists of sub-disciplines. For example, 'Branches of Psychology' becomes 'Psychology'. Similarly, terms in the format 'X by issue' are presented to the user as 'X'.

Screenshots of an example term prominence comparison for part 1 and an example recommendation for part 2 can be seen in Figure 4.8.


Command Prompt - python Compare.py

```
Domain prominence comparison
-----------
Golf has less prominence in your tweets than in your LinkedIn profile.

Is this assessment correct?
Type 'y' for yes or 'n' for no.
If you have no interest in the area at all please type 0.
If this area was of interest to you in the past, but is not anymore, please type
 1.
```

(a) Term prominence comparison

Command Prompt - python Compare.py

```
Recommendations
-----------
Law could be added to your LinkedIn profile.

If you would consider adding this area to your LinkedIn profile, type 0
If you have an interest in this area , but would not add it to your profile, typ
e 1
If you have no interest in this area at all, type 2
If this area was of interest to you in the past, but is not anymore type 3
_
```

(b) Recommendation

Figure 4.8: Example term prominence comparison and example recommendation

4.5.2.3 Metrics for parts 1 and 2

The effectiveness of the Extended WordNet Domains-based and DBpedia category-based representations with regard to both parts of the research question are measured for each participant. Accuracy scores are calculated separately for parts 1 and 2. These accuracy scores are then used to calculate a total, per participant, for each representation. The formulae for Part 1 Accuracy, Part 2 Accuracy and the total score are given below:

$$\begin{gathered}
\text{Part 1 Accuracy} = \frac{\text{number of correct comparisons}}{\text{number of comparisons made}} \cdot 100 \\[2em]
\text{Part 2 Accuracy} = \frac{\text{number of correct recommendations made}}{\text{total number of recommendations made}} \cdot 100 \\[2em]
\text{User Total} = \frac{(\text{Part 1 Accuracy}) + (\text{Part 2 Accuracy})}{2} \cdot 100
\end{gathered} \tag{4.1}$$

However, some terms identified in a user's tweets may not be suitable for a user's LinkedIn profile, even if they were identified correctly. Thus, a total value is calculated that also counts recommendations that were correct but that the user would not add to their LinkedIn profile. This is called *Total (all recommendations)*. The *Part 1 Accuracy*, *Part 2 Accuracy*, *Total* and Total (all recommendations) figures for each participant are aggregated to give a final score for each resource. The formula for calculating these scores is as follows:

$$\text{Resource Score} = \frac{\text{sum of scores for each participant}}{\text{total number of participants}} \tag{4.2}$$

Errors associated with each representation are also measured for each participant. For Part 1, one form of error is naturally the percentage of incorrect term prominence comparisons. This is given by 100 − (Part 1 Accuracy). However, some term prominence comparisons may have been made with terms that were not relevant for the user. This is *Part 1 Term Error.* Term prominence comparisons may also have been made with terms that were relevant for the user in the past, but are not anymore. This is *Part 1 Past Error.* The formulae for these errors are given below:

$$\begin{gathered}
\text{Part 1 Term Error} = \frac{\text{number of prominence comparisons made with incorrect terms}}{\text{total number of prominence comparisons made}} \cdot 100 \\[2em]
\text{Part 1 Past Error} = \frac{\text{number of prominence comparisons made containing past terms}}{\text{total number of prominence comparisons made}} \cdot 100
\end{gathered} \tag{4.3}$$

For Part 2, one error is the percentage of recommended terms that were not relevant for the user at all. This is *Part 2 Error*. Another form of error is the percentage of recommended terms that were relevant for the user, but are not anymore. This is *Part 2 Past Error*. The formulae for these errors are given below:

$$\begin{gathered}\text{Part 2 Error} = \frac{\text{number of irrelevant terms recommended}}{\text{total number of recommendations made}} \cdot 100 \\ \text{Part 2 Past Error} = \frac{\text{number of recommendations made containing past interests}}{\text{total number of recommendations made}} \cdot 100\end{gathered} \tag{4.4}$$

The *Part 1 Term Error*, *Part 1 Past Error*, *Recommendation Error* and *Part 2 Past Error* scores for each participant are aggregated using formula 4.2.

4.5.3 Results

4.5.3.1 The social relevance of each resource

The DMOZ subjects chosen in this experiment are ‘Arts’ and ‘Social Sciences’. The comparison of the DBpedia category hierarchy and the Wordnet Domains hierarchy with respect to these subjects is given in the following subsections.

Arts

Subject area	WordNet Domain parent domain	DBpedia broader categories
Architecture[19]	Applied Science	[‘Applied sciences’, ‘Art media’, ‘Construction’]
Folklore[20]	Ethnology	[‘Fiction’, ‘Oral tradition’, ‘Literature by genre’]
Literature[21]	Humanities	[’Arts’, ‘Humanities’, ‘Writing’, ‘Written communication’]
Mythology[22]	Religion	[‘Folklore’, ‘Oral tradition’, ‘Literature by genre’]
Television[23]	Free Time	[‘Art media’, ‘Entertainment’, ‘Video’, ‘Broadcasting’]
Radio[24]	Free Time	[‘Communications technology’, ‘Wireless’, ‘Broadcasting’]

Table 4.2: The WordNet Domains parent domains and DBpedia broader categories for the ‘Television’, ‘Architecture’ and ‘Literature’ subject areas.

The subjects in the DMOZ ‘Arts’[25] section that have equivalents in WordNet Domains and DBpedia are as follows:

[19] http://dbpedia.org/page/Category:Architecture

[20] http://dbpedia.org/page/Category:Folklore

[21] http://dbpedia.org/page/Category:Literature

[22] http://dbpedia.org/page/Category:Mythology

[23] http://dbpedia.org/page/Category:Television

[24] http://dbpedia.org/page/Category:Radio

[25] http://dmoztools.net/Arts/

[‘Architecture’, ‘Dance’, ‘Literature’, ‘Music’, ‘Myths and Folktales’, ‘Photography’, ‘Radio’, ‘Television’, ‘Theatre’]

The WordNet Domains hierarchy has separate domains for ‘Mythology’ and ‘Folklore’. Conversely, ‘Radio’ and ‘Television’ are represented by a single domain, ‘Radio and TV’.

For three of these subjects, ‘Dance’[26], ‘Music’[27] and ‘Photography’[28], the WordNet Domain super-domain is ‘Arts’. However, SKOS broader categories for each of the DBpedia ‘Dance’, ‘Music’ and ‘Photography’ categories are ‘Performing Arts’, ‘Performing Arts’ and ‘Art media’, respectively. Thus, the DBpedia category hierarchy for these three subjects is also similar to that of DMOZ.

Five subjects in which the WordNet Domains hierarchy differs from the DMOZ hierarchy are ‘Architecture’, ‘Folklore’, ‘Literature’, ‘Mythology’ and ‘Radio and TV’. The parent domains for each of WordNet Domains equivalents of these six areas are listed in Table 4.5.3.1. Also listed in Table 4.5.3.1 are the DBpedia broader categories for the DBpedia category equivalents for these six areas[29]. Note that the DBpedia categories ‘Architecture’, ‘Literature’ and ‘Television’ each have a broader category that relates to the subject of art. Note also that the ‘Television’ and ‘Architecture’ DBpedia categories share a broader category, ‘Art media’. Note also that one of the broader categories for ‘Architecture’ is ‘Applied Sciences’. This is similar to the WordNet Domains super-domain for ‘Architecture’.

[26] `http://dbpedia.org/page/Category:Dance`

[27] `http://dbpedia.org/page/Category:Music`

[28] `http://dbpedia.org/page/Category:Photography`

[29] Note that the list of broader DBpedia categories has been truncated to save space.

Social Sciences

Subject area	WordNet Domain parent domain	DBpedia broader categories
Archaeology[30]	History	[‘Anthropology’, ‘Humanities’, ‘Museology’]
Geography[31]	Earth	[‘Earth sciences’, ‘Social sciences’]
History[32]	Humanities	[‘Humanities’, ‘Past’]
Linguistics[33]	Humanities	[‘Anthropology’, ‘Humanities’, ‘Social sciences’]
Philosophy[34]	Humanities	[‘Abstraction’, ‘Humanities’, ‘Thought’]
Psychology[35]	Humanities	[‘Behavioural sciences’, ‘Mind’]

Table 4.3: The WordNet Domains parent domains and DBpedia broader categories for the ‘Television’, ‘Architecture’ and ‘Literature’ subject areas.

The subcategories in the DMOZ ‘Social Sciences’[36] category that have equivalents in WordNet Domains and DBpedia are as follows:

[‘Anthropology’, ‘Archaeology’, ‘Geography’, ‘History’, ‘Law’, ‘Linguistics’, ‘Philosophy’, ‘Psychology’, ‘Sociology’]

For three of these subjects, ‘Anthropology’, ‘Law’ and ‘Sociology’, the WordNet Domains parent domain is ‘Social Science’. However, ‘Social Sciences’ is a SKOS broader category for each of the ‘Anthopology’[37], ‘Law’[38] and ‘Sociology’[39] DBpedia categories. Once again, while the WordNet Domains hierarchy is similar to the DMOZ hierarchy for these three subjects, so too is the DBpedia category hierarchy.

[30] `http://dbpedia.org/page/Category:Archaeology`
[31] `http://dbpedia.org/page/Category:Geography`
[32] `http://dbpedia.org/page/Category:History`
[33] `http://dbpedia.org/page/Category:Linguistics`
[34] `http://dbpedia.org/page/Category:Philosophy`
[35] `http://dbpedia.org/page/Category:Psychology`
[36] `http://www.dmoz.org/Science/SocialSciences/`
[37] `http://dbpedia.org/page/Category:Anthropology`
[38] `http://dbpedia.org/page/Category:Law`
[39] `http://dbpedia.org/page/Category:Sociology`

Six subjects in which the WordNet Domains hierarchy differs from the above hierarchy are ‘Archaeology’, ‘Geography’, ‘History’, ‘Linguistics’, ‘Philosophy’ and ‘Psychology’. Table 6.4 is an analogue of Table 4.5.3.1 for these domains[40]. Note that two of the DBpedia category equivalents for these categories ‘Geography’ and ‘Linguistics’ have ‘Social sciences’ as their broader categories. Note also that the ‘History’, ‘Linguistics’ and ‘Philosophy’ DBpedia categories have ‘Humanities’ as a broader category, similarly to the corresponding WordNet Domain labels.

Based on Tables 4.2 and 4.3, it appears that the DBpedia category hierarchy has greater similarity with the DMOZ hierarchy than the WordNet Domains hierarchy. This is a consequence of the fact that a DBpedia category can have multiple broader categories, while each WordNet Domain only has one super-domain. Note that this information is not revealed by studies such as Abel et al. [28] and Orlandi et al. [29] that employ external linguistic resources. Such approaches do not examine how the hierarchy of their chosen linguistic resource relates to personalisation performance.

[40] Similarly to Table 4.5.3.1, the broader categories in Table 4.2 are truncated.

4.5.3.2 Accuracy and Error for parts 1 and 2 of the research question

The number of questions asked for each participant for each part of the research question is given in Table 4.4. Note that for participant 5 no comparisons were made for the DBpedia approach. This participant had a sparse LinkedIn profile. The accuracy scores for each part of the research question are given in Table 4.5. The error percentages are given in Table 4.6.

Participant	Part 1 DBC Counts	Part 1 EWND Counts	Part 2 DBC Counts	Part 2 EWND Counts
1	4	1	1	1
2	2	7	1	4
3	21	48	7	2
4	5	14	5	10
5	0	4	2	3
6	6	11	11	4
7	7	18	6	3
8	5	17	6	3

Table 4.4: Part 1 and Part 2 Accuracy percentages. 'EWND' denotes the Extended WordNet Domains-based representation and 'DBC' denotes the DBpedia category-based representation.

	Part 1 Accuracy	Part 2 Accuracy	Total	Total (all recommendations)
EWND	39.80%	29.58%	34.69%	54.37%
DBC	61.94%	51.34%	56.64%	69.95%

Table 4.5: Part 1 and Part 2 Accuracy percentages. 'EWND' denotes the Extended WordNet Domains-based representation and 'DBC' denotes the DBpedia category-based representation.

	Part 1 Term Error	Part 2 Error	Part 1 Past Error	Part 2 Past Error
EWND	29.06%	29.79%	3.70%	1.00%
DBC	16.22%	14.29%	0	0

Table 4.6: Error rate percentages. 'EWND' denotes the Extended WordNet Domains-based representation and 'DBC' denotes the DBpedia category-based representation.

4.6 Discussion

Notice that, for all but one of the participants, more comparisons are performed using Extended WordNet Domains than DBpedia categories. This is due to the total number of labels in each resource. Extended WordNet Domains has a total of labels is 170, whereas there are more than one million DBpedia categories. This means that it is far more likely for terms to be shared between a user's Twitter and LinkedIn term lists when using Extended WordNet Domain labels than when using DBpedia categories. A key implication of this disparity concerns the issue of coverage. For example, if Extended WordNet Domains yielded a higher absolute number of correct comparisons than DBpedia categories it could be argued that the former covers more user interests and knowledge. However, assessing coverage would require a complete list of user interests and knowledge for each participant for each resource. The process required to create such a list is outside the scope of this research.

Neither resource consistently provides more recommendations than the other. This is because recommendations are performed on the basis of relative ranks between term lists, which does not depend on the total number of labels in each resource.

The Extended WordNet Domains-based representation shows a 10% greater number of incorrectly identified terms than the DBpedia category-based representation. It also shows more than twice the percentage of incorrect recommendations. A reason for this can be found in the property that appears to give DBpedia greater social relevance than WordNet Domains i.e. the looser inheritance hierarchy of the former. Each representation is generated by identifying labels in text and obtaining the parents of these labels. The fact that each WordNet Domain only has a single parent creates a requirement of perfect accuracy i.e. that single parent must accurately describe the user's information. The fact that each DBpedia category has multiple parents means that only one or more of these parents must accurately describe the user's information.

The marked difference between the 'Total' and 'Total (all recs.)' columns in Table 4.2 is also noteworthy. This suggests that there are certain subjects the participants intended for Twitter, but not for LinkedIn. However, there is an additional conclusion to be drawn. The difference between the 'Total' and 'Total (all recommendations)' columns is 17%. The figures in the 'Total' column are obtained by adding the 'Part 1 Accuracy' and 'Part 2 Accuracy' figures and dividing by two. The figures in the 'Total (all recommendations)' are obtained by changing the 'Part 2 Accuracy' figure only. That is, the 'Part 1 Accuracy' figure is the same in the calculation of the 'Total (all recommendations)' column as it is in the calculation of the 'Total' column. Thus, the 17% difference between the totals represents a 34% difference in the recommendation scores. This means that the accuracy of the recommendations for the DBpedia-based representation was 84%. This is higher than the 80% figure obtained in by Yu et al. [104] who make term recommendations for users based on social media content they have generated.

However, the discrepancy between the 'Total' and 'Total (all recs.) columns merits further investigation. The fact that LinkedIn is a professional network suggests that terms participants specifically chose not to include on their LinkedIn profile do not relate to their professional activities. Recall from Section 2.6 that Heitmann et al. [37] model the Knowledge characteristic in the context of a user's profession. Furthermore, based on a review of User Modeling approaches, Plumbaum et al. associate the Knowledge characteristic with a user's profession [105]. This suggests that the terms users specifically chose not to include on their LinkedIn profile do not relate to the Knowledge characteristic, but to the Interests characteristic. This raises the question of how to automatically determine whether an identified term relates to the Interests or Knowledge characteristic.

One aspect of this experiment in need of improvement is the procedure for part 1 of the research question. During this part of the experiment, some participants said that they could not be sure about the relative weights of individual terms in their tweets and LinkedIn profile. However, in this case users were instructed to answer negatively so as not to artificially inflate scores. One way of overcoming this problem could be to leverage the intuition behind the Kendall Tau coefficient, which measures the level of agreement between two different rankings of the same list of items [106]. Central to the calculation of Kendall Tau are the *concordance* and *discordance* measures between pairs of terms. Concordance occurs when the relative positions of two terms is the same in each ranking i.e. the same term is higher in both rankings. Discordance occurs when the relative positions of two terms is different in each ranking i.e. a term is higher in one ranking and lower in another. Consider the sample ranked lists in Figure 4.9. The pair `['physics', 'biology']` is concordant between the lists because 'physics' is ranked higher than 'biology' in each list. The pair `['politics', 'sport']` is discordant between the lists because 'politics' is ranked higher than 'sport' in the Twitter term-list but lower in the LinkedIn term-list.

To apply the principle of concordance and discordance for part 1, the Twitter and LinkedIn term-lists would first be generated using the process described in Section 4.4. Concordant and discordant pairs would be identified. The participant would then be shown each of the Twitter

```
Twitter term-list:  [‘politics’, ‘physics’, ‘sport’,
‘linguistics’ , ‘biology’]


LinkedIn term-list:  [‘sport’, ‘politics’, ‘physics’,
‘biology’, ‘linguistics’]
```

Figure 4.9: Two different rankings of a set of 5 terms.

and LinkedIn term-lists separately. For each term-list, the participant would identify terms that should be higher or lower in the ranking. After the participant had completed this process a new ranking will have been created for each of the Twitter and LinkedIn term lists. Concordant and discordant pairs would then be identified for the new rankings. The sets of concordant and discordant pairs would be compared in order to identify differences between the two. Difference would be defined as a pair that had originally been concordant now being discordant and vice versa. Suppose that, for the DBpedia category-based representation, after the participant had re-ranked the term-lists in Figure 8 the formerly concordant pair `[‘physics’, ‘biology’]` was now discordant. This would mean that the relative prominence of these terms *should* have been different between the Twitter and LinkedIn term-lists, and that the DBpedia category-based representation failed to represent this fact.

4.7 Conclusion

This chapter described a comparison between two external linguistic resources: a purely lexical resource (WordNet) and a general knowledge base (DBpedia). The comparison took the form of an investigation of the ways in which users represent their interests and knowledge through their LinkedIn profile with the way they represent the same characteristics through their tweets. The results of this experiment indicate that DBpedia’s greater social relevance appears makes it more suitable for this task than Extended WordNet Domains.

However, the results of this experiment also revealed that a representation being able to accurately identify that a list of terms is relevant for a user is not enough. The representation must accurately reflect the distinction between different categories. An investigation of how to determine a structured representation’s ability to accurately make this distinction is the subject of the next chapter. This investigation will examine this ability in the context of multiple domain-distinct recommendation tasks involving a variety of user models and recommendation methods. This investigation will assess the ability of different structured representations to accurately represent category distinctions, and how this ability relates to these structured representations' ability to provide accurate recommendations.

Chapter 5

Investigating the influence of content domain

The experiment described in the previous chapter indicated that the DBpedia category-based representation can more accurately represent user interests and knowledge than the Extended WordNet Domains-based representation. However, accuracy alone is not enough: a structured representation must also be able to distinguish between categories. This was evidenced by the fact that participants indicated that numerous recommendations made from their tweets were correct, but were unsuitable for their LinkedIn profile. The results of the experiment described in the previous chapter show that if this distinction is not respected, recommendation accuracy can suffer. The experiment described in this chapter investigates this distinction through multiple recommendation tasks. These tasks vary with respect to content domain, user model and recommendation method. The sources of user information employed are tweets from personal Twitter accounts, publications users have written[1] and content users have liked.

This experiment will investigate the extent to which a structured representation accurately reflects category distinctions, and how this affects the ability of that structured representation to facilitate effective recommendation. In order to fully investigate the nature of this relationship, different user modeling approaches and recommendation methods will be employed. For example, it may be the case that a structured representation that does not accurately reflect domain category distinctions will not facilitate effective recommendation irrespective of the user modeling approach or recommendation method that is applied. On the other hand, a structured representation that does not accurately reflect domain category distinctions may facilitate effective recommendation if a particular user model and/or recommendation method is applied.

[1] Due to a change in the LinkedIn API (described here: https://developer.linkedin.com/support/developer-program-transition), the required LinkedIn profile information can no longer be collected.

Section 5.1 describes the objective of this experiment. **Section 5.2** describes the case that will be investigated in this experiment. **Section 5.3** describes the question of this experiment. **Section 5.4** describes the process for generating the structured representations used in this experiment. **Section 5.5** describes the experiment design. **Section 5.6** discusses the results of the experiment and **Section 5.7** concludes.

5.1 Objective of this experiment

Recall that in Section 4.6, a question was raised concerning how to filter a structured representation so that it only contains information relating to a single user characteristic (i.e. interests or knowledge). Recall also that the third objective of the Research Question of this thesis is to investigate the influence of content domain on the efficacy of structured representations in personalisation tasks. The objective of the experiment described in this chapter is thus to investigate the relationship between a structured representation's ability to accurately reflect domain category distinctions and that representation's ability to facilitate effective recommendation.

5.2 Case description

Researchers such as Gao et al. [57] and Ma et al. [76] argue that aggregating user information from multiple sources generates user models of greater breadth and depth than using only one source. Such approaches to profile aggregation employ an additive approach. For example, if source A has two mentions for 'Linguistics' and Source B has three mentions for 'Linguistics', the aggregated profile has five mentions for 'Linguistics'. Alternatively, a specific weight can be given to one source. For example, Source B may be given twice as much weight as Source A. This approach will henceforth be referred to as *Additive-Aggregation.* There are two issues to be considered in the Additive-Aggregation approach:

1. The ranks of common interests expressed on multiple sources can differ. For example, a user may represent 'Linguistics' as the most prominent subject on one source but as the third most prominent subject on another source. Thus, when performing profile aggregation, it may be necessary to use the relative ranks of terms rather than summing across different profiles. This approach will henceforth be referred to as *Rank-Aggregation.*

2. There are certain subjects that users purposely do not express through certain sources. For example, a user may be interested in physics but decide not to discuss it on their LinkedIn profile. An aggregation strategy that does not account for this fact may hinder

personalisation for this user in the professional domain. For example, in a job recommendation approach, this aggregated profile could cause irrelevant jobs to be recommended. Thus, when performing multi-domain personalisation, it may be necessary to filter the aggregated profile according to domain.

Furthermore, it may be the case that the user modeling method itself is not the strongest determinant of recommendation performance. It may be the case that, if the structured representation employed does not respect the category distinctions of the content being analysed, no user model will yield accurate recommendations.

The experiment described in this chapter examines the above issues by means of multiple recommendation tasks spanning different domains, user models and recommendation approaches.

The author argues that this case meets the requirements of Case Study 3 described in Section 2.6.5 for the following reasons:

- It concerns content from a variety of domains.
- By employing different user models, this experiment will provide an in-depth examination of the relationship between the accurate reflection of domain category distinctions and recommendation performance. For example, it may be the case that no user model can provide accurate recommendations if the underlying structured representation does not reflect domain category distinctions. On the other hand, it may be the case that a particular user model may be able to provide accurate recommendations, even if domain category distinctions are not reflected.

5.3 Question of this experiment

The research question of this experiment is as follows:

How does the ability of a structured representation to accurately reflect domain category distinctions affect the accuracy of recommendations provided using that representation?

This question will be examined with respect to two recommendation tasks:

1. A task that analyses content users have created in order to recommend items for those users. Recommended content will be from the news and jobs domains.

2. A task that analyses items users have liked in order to recommend new items for those users. Recommended content will be from the music and movie domains.

Task 1 will use information from users' tweets and abstracts they have written. Note that studies such as Abel et al. [7], Kapanipathi et al. [22] and Mendes et al. [59] analyse a user's tweets alone in order to represent that user's interests. Similarly, approaches such as Zeng et al. [107], Wu et al. [108] and Taylor and Richards [109] analyse solely publications a user has authored in order to represent that user's knowledge. Thus, Task 1 also allows for an investigation of the Interests/Knowledge distinction raised in Section 4.6.

5.4 Generating the structured representation

Since Task 1 will investigate the issue raised in Chapter 4, the analysis method employed in Task 1 is the same as that employed in Chapter 4. However, to ensure that the experiment described in this chapter is not biased towards a single method of generating structured representations, two external analysis methods are employed in Task 2. These are Babelfy[2] and AlchemyAPI[3]. Each of these tools maps terms in text to DBpedia resources. Each tool assigns a real-valued relevance score between 0 and 1 to each identified DBpedia resource, with 1 being the highest. This relevance score indicates the significance of this resource for the text.

Recall from Section 3.4 that AlchemyAPI has been employed in a variety of personalisation tasks. Babelfy is selected for its state-of-the-art performance. In [110], the creators of Babelfy compare Babelfy to multiple state-of-the-art approaches in disambiguation and entity linking[4] on six different datasets and find that Babelfy obtains the most robust performance across the different datasets.

5.4.1 Creating the different user models for Task 1

Recall that the sources of user information for this task are users' tweets and research abstracts. Recall also that, as well as investigating the research question of the study described in this chapter, this task will also investigate the Interests/Knowledge distinction raised in Section 4.6. For this reason, multiple user models are generated in this task.

[2] `http://babelfy.org/`

[3] `http://www.alchemyapi.com/`

[4] Entity linking is the task of identifying a mention of an entity, when the mention does not contain that entity. For example, in the sentence 'She is the Chancellor of Germany', 'She' refers to Angela Merkel.

The following user models are generated:

1. Publication User Model - Terms identified in the user's research publications.
2. Twitter User Model - Terms identified in the user's tweets.
3. Additive-aggregated User Model - Terms identified in both profiles, and combined using the additive method.
4. Abstract-only User Model - Terms that were identified in the user's research publications but not their tweets.
5. Twitter-only User Model - Terms that were identified in the user's tweets but not in their research publications.
6. Rank-aggregated User Model - Terms identified in both profiles, rank-aggregated.

The processes by which each of these User Models are generated are described in the following subsections.

5.4.1.1 Twitter User Model and Publication User Model

These User Models consist of term-lists generated using the DBpedia category-based representation described in Section 4.4.

5.4.1.2 Aggregated User Models

The first step in creating each aggregated User Model is to generate separate term-frequency lists for a user's research publications and tweets. This is performed using the DBpedia-category approach described in Section 4.4, except that the weight calculation step is not performed. The Additive-Aggregated User Model is generated by summing the Publication and Twitter term-lists. The process for generating the Rank-Aggregated User Model is described below.

Rank-aggregated User Model
The Schulze rank-aggregation method is employed to generate the Rank-Aggregated User Model. For a description of the Schulze method, please see Appendix C.

Are the Publication and Twitter term-lists strict weak orders?
Recall from Appendix C that the Schulze method requires strict weak orders as input. This raises the question of whether or not the Publication and Twitter term-lists generated by the

analysis described in this chapter fit the criteria of a strict weak order. Recall that a strict weak order has the following properties for all element pairs a and b:

- Asymmetricity - This property is satisfied if only one of the following statements is true:
  - $a \succ b$ (i.e. a is ranked higher than b)
  - $b \succ a$ (i.e. b is ranked higher than a)
  - $a \approx b$ (i.e. neither $a \succ b$ nor $b \succ a$)
- Irreflexivity - This property states that $a \approx a$. In other words, a cannot be ranked higher or lower than itself.
- Transitivity - This property states that if $a \succ b$ and $b \succ c$, then $a \succ c$. That is, if a is ranked higher than b, and b is ranked higher than c, then a must be ranked higher than c.
- Negative Transitivity - This property states that if $a \not\succ b$ and $b \not\succ c$ then $a \not\succ$c. That is, if a is not ranked higher (i.e. is ranked lower or equal) than b, and b is not ranked higher than c, then a cannot be ranked higher than c.

Note that the relative rank of two terms a and b depends on the number of times a and b have been mentioned.

Only one of the following statements can be true: (i) a has more mentions than b (ii) b has more mentions than a (iii) a and b have the same number of mentions. Thus, the Publication and Twitter ranked term-lists display the asymmetricity property.

A term a cannot have higher or lower mentions than itself. Thus, the Publication and Twitter term-lists display the irreflexivity property.

If a has more mentions than b and b has more mentions than c, then a must have more mentions than c. Thus, the Publication and Twitter term-lists display the transitivity property.

If a has less than or equal the number of mentions of b, and b has less than or equal the number of mentions of c, then a must have less than or equal the number of mentions of c. Thus, the Publication and Twitter term-lists display the negative transitivity property.

Therefore, the Publication and Twitter term-lists are strict partial orders.

The difference between Additive-Aggregation and Rank-Aggregation

Consider the term-frequency lists in Figure 5.1. The numeric values represent the total number of mentions received by each term. Figure 5.2 shows the Additive-Aggregated User Model obtained from these two term-lists. The numbers in parentheses besides each term denote that term's total number of mentions. Figure 5.3 shows the Rank-Aggregated User Model obtained from the term-frequency lists in Figure 5.1. Note that in the Additive-Aggregated User Model, 'Term3' is ranked higher than 'Term2'. This is because the *absolute* number of mentions for 'Term3' is higher than for 'Term2'. However, in the Rank-Aggregated User Model, 'Term2' and 'Term3' are tied. This is because 'Term2' appears higher than 'Term3' in one list and 'Term3' appears higher than 'Term2' in the other.

```
[‘Term1’:20,‘Term3’:18,‘Term2’:8,‘Term4’:7]



[‘Term1’:5,‘Term2’:4,‘Term3’:3,‘Term4’:2]
```

Figure 5.1: Two different term-frequency lists for a collection of 4 terms.

```
[‘Term1’ (20),‘Term3’ (18),‘Term2’ (8),‘Term4’ (7)]
```

Figure 5.2: The Additive-Aggregated User Model generated from the term-frequency lists in Figure 5.2.

```
[‘Term1’,(‘Term3’ ,‘Term2’),‘Term4’]
```

Figure 5.3: The Rank-Aggregated User Model generated from the term-frequency lists in Figure 5.2. ‘Term 3’ and ‘Term2’ are tied.

	Resolvability	Pareto	Reversal Symmetry	Monotonicity	Independence of clones	Smith	Smith-IIA	Condorcet	Condorcet Loser	Majority For Solid Coalitions	Majority	Majority Loser	Participation	MinMax Set	Prudence	Polynomial Runtime
Baldwin	Y	Y	N	N	N	Y	N	Y	Y	Y	Y	Y	N	N	N	Y
Black	Y	Y	Y	Y	N	N	N	Y	Y	N	Y	Y	N	N	N	Y
Borda	Y	Y	Y	Y	N	N	N	N	Y	N	N	Y	Y	N	N	Y
Bucklin	Y	Y	N	Y	N	N	N	N	N	Y	Y	Y	N	N	N	Y
Copeland	N	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	N	N	N	Y
Dodgson	Y	Y	N	N	N	N	N	Y	N	N	Y	N	N	N	N	N
Instant Runoff	Y	Y	N	N	Y	N	N	N	Y	Y	Y	Y	N	N	N	Y
Kemeny-Young	Y	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	N	N	N	N
Nanson	Y	Y	Y	N	N	Y	N	Y	Y	Y	Y	Y	N	N	N	Y
Plurality	Y	Y	N	Y	N	N	N	N	N	N	Y	N	Y	N	N	Y
ranked pairs	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	N	Y	Y
Schulze	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	Y	Y	Y
Simpson-Kramer	Y	Y	N	Y	N	N	N	Y	N	N	Y	N	N	N	Y	Y
Slater	N	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	N	N	N	N
Young	Y	Y	N	Y	N	N	N	Y	N	N	Y	N	N	N	N	N

Table 5.1: Comparison of the Schulze Method with other voting methods [111].

The suitability of the Schulze method for the case study described in this chapter Schulze provides a comparison of the Schulze method with other aggregation methods [111]. This is displayed in Table 5.1. Some of the criteria used in this comparison are not relevant for the study described in this chapter, while other criteria are relevant. The irrelevant and relevant criteria are described in the following subsections.

Irrelevant criteria for the study described in this chapter

Resolvability

Methods that satisfy this criterion have a high likelihood of picking a unique winner. This is naturally important in elections, as only one candidate can win. However, in ranking terms that represent user interests and knowledge, there need not necessarily be an outright winner.

Reversal Symmetry

This criterion states that if the ranking of items is reversed, the aggregated ranking should also be reversed. In other words, in cases where voters provide rankings of their favourite candidates and least-favourite, the same candidate should not top both lists. This criterion is not relevant for the study described in chapter as no list representing users' disinterest or lack of knowledge is generated.

Monotonicity

This criterion states that raising the rank of a candidate a should never hurt a's chances of winning. Once the User Models have been generated in the study described in this chapter, they will not be changed. Thus, the Monotonicity criterion is not relevant for the study described in this chapter.

Independence of clones

This criterion states that running a large number of similar alternatives (clones) should not affect the aggregated ranking. This criterion measures a method's ability to account for the fact that similar candidates may take votes from each other. Schulze provides the following example:

"In 1969, when the Canadian city that is now known as Thunder Bay was amalgamating, there was some controversy over what the name should be. In opinion polls, a majority of the voters preferred the name The Lakehead to the name Thunder Bay. But when the polls opened, there were three names on the referendum ballot: Thunder Bay, Lakehead, and The Lakehead. As the ballots were counted using plurality voting, it was not a surprise when Thunder Bay won. The votes were as follows: Thunder Bay 15870, Lakehead 15302, The Lakehead 8377."

Such issues are not a concern for the study described in this chapter.

Smith-IIA (Independence of Irrelevant Alternatives)
Consider a division of the entire set A of items into two subsets, B and C such that for every candidate b in B and every candidate c in C, $b \succ c$. B is known as the *Smith Set*. The Smith-IIA criterion states that adding or removing a candidate from C should not change the winner of the election. The Smith-IIA criterion is not relevant for the study described in this chapter, as no terms will be added to or removed from the User Models after they have been generated.

Participation
Participation states that adding a set of ballots in which a wins a pairwise comparison with b should not change the winner from a to b. The Schulze Method does not satisfy this criterion. However, once the Rank-Aggregated User Model has been generated, no changes will be made to it. Thus, the Participation criterion is not relevant for the study described in this chapter. Note that this is the only criterion not satisfied by the Schulze Method.

Polynomial runtime
This criterion states that the maximum length of time required by a method to run is represented by a polynomial expression in the number of ballots. Since the number of ballots will always be 2, runtime considerations are not a factor in the study described in this chapter.

Relevant criteria for the study described in this chapter

Pareto
This criterion states that if all voters prefer a candidate a to a candidate b, b should not be the winner.

Smith
The Smith criterion states that the winning candidate is always chosen from the Smith Set, described in the preceding subsection..

Condorcet
A *Condorcet Winner* is a candidate who wins a pairwise comparison against every other candidate. A voting method satisfies the Condorcet criterion if it always selects the Condorcet Winner, when one exists.

Condorcet Loser

A *Condorcet Loser* is a candidate who loses a pairwise comparison against every other candidate. A voting method satisfies the Condorcet Loser criterion if it never selects the Condorcet Loser.

Majority for solid coalitions

Suppose there is a subset B of A consisting of candidates who are preferred by all voters to every candidate outside of B. A voting method satisfies the 'Majority for solid coalitions' criterion if it always selects the winner from B.

Majority

If more than half the voters rank a candidate a over every other candidate, then a should win. Note that in the study described in this chapter, only two User Models are being aggregated. Thus, 'over half' means 'all' in this study, meaning the Majority condition is covered by the Condorcet criterion.

Majority loser

If more than half the voters rank every other candidate over a, a must not win. Similarly to the Majority criterion, the Majority Loser criterion is covered by the Condorcet Loser criterion.

MinMax Set

The *MinMax Score* of a subset B is the strength of the strongest pairwise victory of a candidate outside of B against a candidate from B. The Schulze Method always chooses a winner from the set whose MinMax Score is lowest. In other words, the winner is always chosen from the set whose worst candidate is better than the worst candidate in any other set.

Prudence

Recall from Appendix C the definition of a path between two candidates a and b in the pairwise ranking matrix N. The strength of this path is denoted *(N[a,b],N[b,a])*. A cycle is a path in N that starts and ends at the same cell. Cycles occur when ballots contain candidates who are given the same preference e.g. two candidates are ranked second. The strength of the cycle is the strength of the weakest link on it. Let *max* equal the maximum of all the cycle strengths in N. A voting method is prudent if:

- Given any two items a and b, if *(N[a,b],N[b,a])* is greater than *max* then *N[a,b]* is greater than *N[b,a]*.
- Given any two items a and b, if *(N[a,b],N[b,a])* is greater than *max* then b cannot win.

If cycles exist in N, the Prudence criterion minimises the size of the largest cycle.

Note from Table 5.1 that the Schulze method is alone in satisfying each of the requirements listed in this section. It is for this reason that the Schulze method is chosen for the study described in this chapter.

The Schulze algorithm is implemented using the Python Vote Core package[5]. This library was chosen as it is referenced in a W3C discussion on Voting Systems[6].

5.4.1.3 Twitter-only User Model and Abstract-only User Model

Section 4.5.2.1 discusses a recommendation method based on ranks. This method uses the term-list generated from one source (LinkedIn profile information) as a basis for recommending terms from a term-list generated using another (tweets). The recommendations generated using this method were of high accuracy. In the study described in this chapter, this recommendation method is employed in order to generate source-specific User Models. The first step in generating source-specific User Models is to compile lists of terms that appear only in the user's tweets or only in their research publications. The Rank-Aggregated User Model is then used as a basis for identifying terms in these lists whose rank is high enough to be included in the Twitter-only/Abstract-only User Model. For example, given the Rank-Aggregated User Model in Figure 5.3, the Abstract-only User Model would consist of the terms appearing only in the user's publications that were of rank 3 or higher. The Twitter-only User Model is generated analogously. The Rank-Aggregated User Model is used as it is created by an aggregation process that specifically accounts for the ranking of each source. This is required since the recommendation method employed depends on ranks.

5.5 Experiment Design

5.5.1 Dataset description

5.5.1.1 Task 1

As in the previous chapter, the user's 1000 most recent tweets are collected. Research publications the user has written are collected (with associated keywords, if present), with a minimum of 5 and a maximum of 20. The DBLP database[7] is used to identify a participant's

[5] https://github.com/bradbeattie/python-vote-core

[6] Available here: https://lists.w3.org/Archives/Public/public-webizen/2014May/0000.html

[7] http://dblp.uni-trier.de/

publications. The lower end of this range was chosen with reference to the study performed by Zeng et al in which user research interests are modelled through research publications they have authored [107]. The upper end of this range was decided with reference to the DBLP list of publications per author[8]. In this list, the number of authors with 20 publications or less comprises slightly more than 95% of the total number of authors. A graph displaying this distribution can be seen in Figure 5.4. Publication abstracts, keyword lists and tweets are converted to term-lists using the analysis process described in Section 4.4. These term-lists are then used to create the six user models being investigated in the case study described in this chapter.

The news article corpus created by Greene and Cunningham is used [112][9]. This dataset is used as it was specifically designed to contain a variety of topics, having been created for the purposes of evaluating of a classification algorithm. Its articles are divided into the following topics: Business, Entertainment, Politics, Sport, Technology. For academic vacancies, the vacancy list of the University of Oxford is used[10]. 200 items are chosen at random from each of these corpora. For the news corpus, 40 articles are chosen randomly from each topic. The full list of categories for these vacancies is given in Appendix E.

News articles and academic vacancies are converted to term-lists using the analysis process described in Section 4.4. Links are discarded for reasons similar to those given in Section 3.4. For example, an academic vacancy may contain a link to the homepage of the relevant department. Analysis of this link could introduce terms related to every research area being pursued by this department.

5.5.1.2 Task 2

For the movie recommendation task, the MovieLens dataset created by Harper and Konstan is used [113]. The MovieLens dataset is a collection of user ratings of movies. Each movie is rated out of 5. A rating of 1 or 2 indicates that the user did not like the movie, while a rating of 3-5 indicates the user liked the movie. The MovieLens dataset also provides genre information for each movie. The full list of MovieLens genres can be found in Appendix F. The dataset has numerous versions, each containing a different number of ratings. The 100K version is used for this follow-up experiment. This dataset consists of 100,000 ratings from 943 users on 1682 movies. For the purposes of this follow-up experiment, only movies rated 3 or higher are required. This is because the experiment described in this chapter

[8] Available here: http://dblp.uni-trier.de/statistics/numberofpublicationperauthor

[9] Available here: http://mlg.ucd.ie/datasets/bbc.html

[10] Available here: https://source.data.ox.ac.uk/archive/public/vacancies-current/latest.rdf

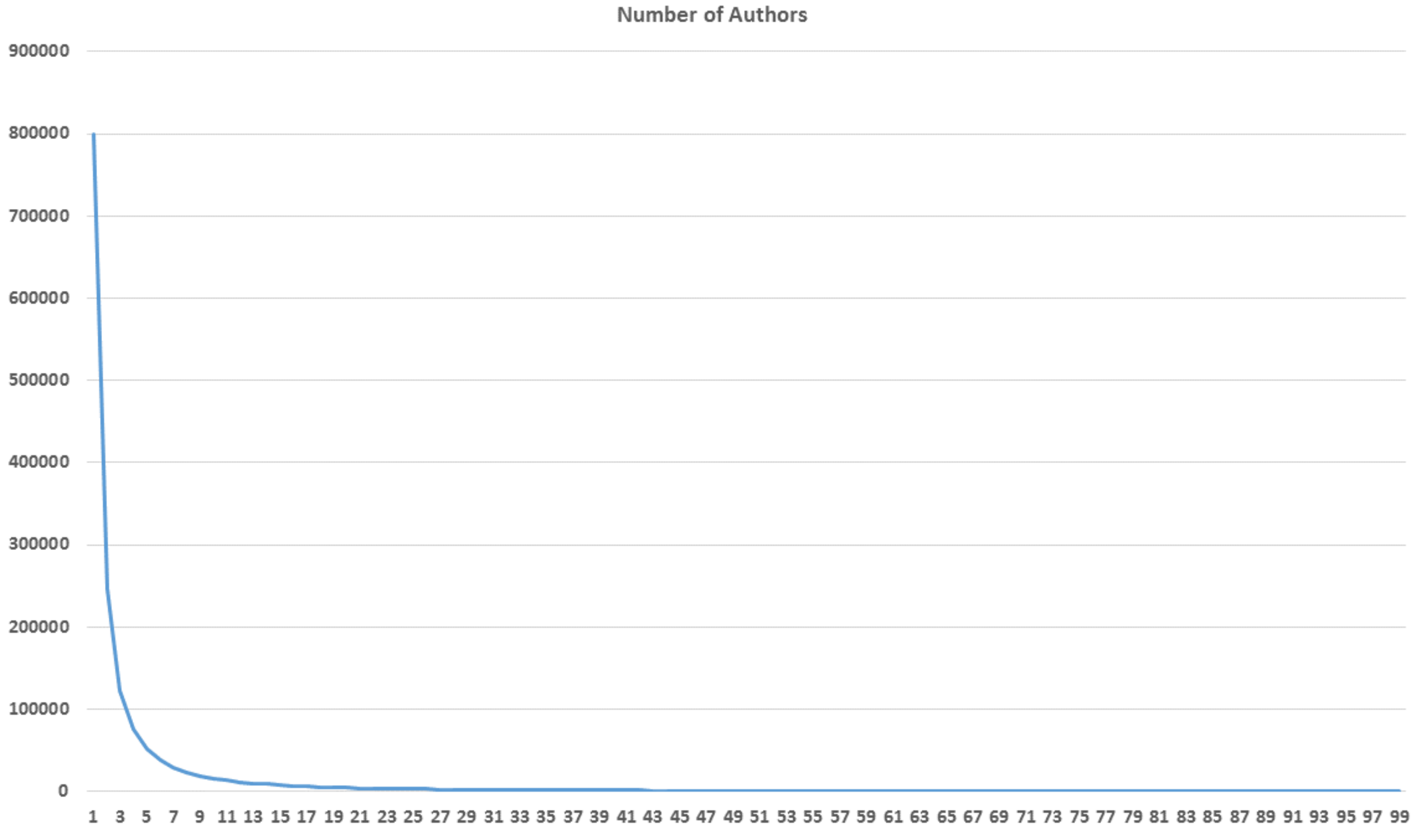


Figure 5.4: A graph of number of authors per number of publications on DBLP.

is only concerned with identifying items a user liked, not items they did not like. Thus, all user ratings of 2 or 1 are filtered out of the dataset. After this filtering, any users with less than 20 ratings in the dataset are removed. This limit is equal to the rating limit imposed by Harper and Konstan in creating the MovieLens dataset. This filtering left 82520 ratings from 845 users. These ratings were then randomly partitioned into five training and test sets. Each user's training set contains 80% of the user's ratings while each user's test set contains the remaining 20% of their ratings. This train/test split is also the same as that of Harper and Konstan. Textual descriptions of the movies in the dataset were obtained from IMDB[11], or Rotten Tomatoes[12], if there was no description for the film on IMDB.

For the music recommendation task, the music dataset created by Oramas et al. is used [114]. This dataset consists of a list of users and the songs that are relevant for these users. The listening information is obtained from Last.fm[13]. Oramas et al. determine relevant songs for each user by calculating an average listening count for each user. A song is relevant for a user if the user's listening count for that song is above the user's average listening count. The dataset also contains tags and a textual description for each song from the site Songfacts.com[14]. In total, the dataset contains 751531 downloads from 5199 users of 8640 items. An analysis of this entire dataset would be computationally intractable. Therefore, a subset of this dataset is created. The procedure for creating this dataset is as follows:

1. The songs are ranked, from highest to lowest, according to the number of users to which the songs are relevant.

2. The top 1500 songs are identified.

3. The user ratings for these 1500 songs are collected.

4. Any user in the resulting ratings list who has less than 20 ratings is removed from the list.

These ratings were then partitioned into five training and test sets, analogously to the MovieLens dataset. In order to categorise this dataset the tags associated with each song are leveraged. The tags are first ordered with respect to the number of times they have been used. Tags are then selected with reference to the Wikipedia list of song styles[15]. The full list of

[11] http://www.imdb.com/

[12] https://www.rottentomatoes.com/

[13] http://www.last.fm/

[14] http://www.songfacts.com/

[15] https://en.wikipedia.org/wiki/List_of_music_styles

genres can be seen in Appendix G.

However, there were certain movies and music descriptions for which each analysis tool was unable to identify any subjects. These items were removed. After this removal, any user who had less than 20 ratings was removed from analysis. The final number of users, ratings and files for the music and movie datasets when analysed by AlchemyAPI can be seen in Table 5.2. The same statistics for Babelfy can be seen in Table 5.3.

Dataset	**Number of users**	**Number of ratings**	**Number of files**
Music	1000	81586	1438
Movies	845	80933	1676

Table 5.2: Movie and Music dataset statistics for AlchemyAPI.

Dataset	**Number of users**	**Number of ratings**	**Number of files**
Music	1000	81572	1441
Movies	844	80906	1677

Table 5.3: Movie and Music dataset statistics for Babelfy.

5.5.2 Generating recommendations

Cosine Similarity is used in each task to measure similarity between items. Cosine similarity applies the *Vector Space Model*, in which a weighted term -list containing n terms is treated as an n-dimensional vector, where each term represents a single dimension [115]. A vector is " 'something' that has magnitude and direction" [116]. Figure 5.4 shows an example weighted term-list and its associated vector. In the vector space model, the similarity between two term-lists can be measured as the Cosine of the angle between their associated vectors. The Cosine Similarity between two n-dimensional vectors $\vec{v1}$ and $\vec{v2}$ is calculated using the following formula:

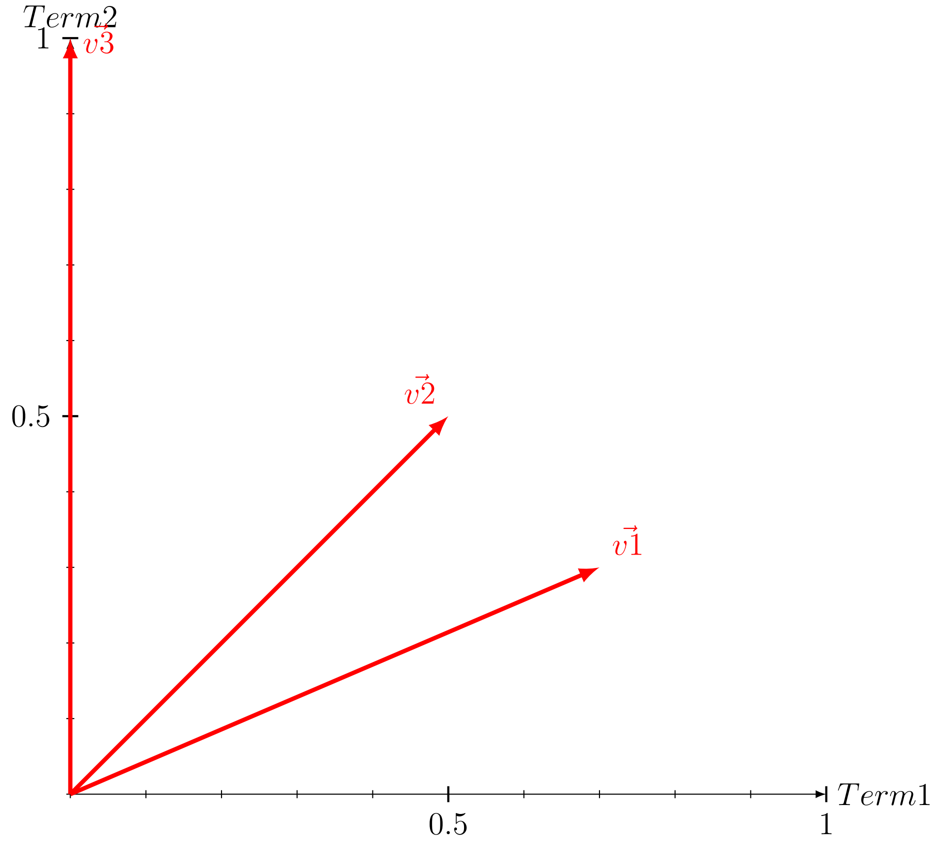
$$sim(\vec{v1}, \vec{v2}) = \frac{\sum_{i=1}^{n} w_{i,v1} \cdot w_{i,v2}}{\sqrt{\sum_{i=1}^{n} w_{i,v1}^2} \cdot \sqrt{\sum_{i=1}^{t} w_{i,v2}^2}} \tag{5.1}$$

So, for the term-lists in Figure 5.5, the Cosine Similarity is calculated as:

$$sim(\vec{v1}, \vec{v2}) = \frac{(0.3 \cdot 0.5) + (0.7 \cdot 0.5)}{\sqrt{(0.3^2 + 0.7^2)} \cdot \sqrt{(0.5^2 + 0.5^2)}} = 0.928 \tag{5.2}$$

It will sometimes be the case that, when comparing a User Model against a term-list representing an individual document, one vector has terms that the other vector does not have. When this occurs, the number of dimensions in the vector space is the sum of the number of unique terms of both vectors. The weight for any feature a vector does not contain is set to 0. This can be seen in vector $\vec{v3}$ in Figure 5.5.

The Cosine similarity measure is frequently used in recommendation tasks, for example academic papers [117], news articles [7], movies [118] and the recommendation of tags on the tagging site Delicious [104].



$\vec{v1}$: `[‘Term1’:0.3, ‘Term2’:0.7]`

$\vec{v2}$: `[‘Term1’:0.5, ‘Term2’:0.5]`

$\vec{v3}$: `[‘Term1’:0, ‘Term2’:1.0]`

Figure 5.5: Sample weighted term-lists and the associated vectors.

5.5.2.1 Task 1

Each User Model is compared with all of the documents in the news/jobs dataset. The 10 items with the highest Cosine Similarity are recommended to the user. However, each term in the term-list representation of an item must have a numerical weight in order to apply the Cosine Similarity measure. The Schulze algorithm does not provide weights for the items generated in the aggregated ranking. Thus, in order to apply the Cosine Similarity measure on the Rank-Aggregated User Model, a means of eliciting weights for the terms contained in this User Model must be identified. Weight elicitation techniques can be divided into two broad categories: (i) Direct - where weights are represented using numerical values (ii) Indirect - where indirect ordinal or interval judgements are used to derive a ratio scale [119]. Since the Cosine Similarity measure works on requires numerical values, direct weight elicitation techniques are appropriate for the study described in this chapter.

Barron and Barrett find that the Rank-Order Centroid (ROC) method is the most effective means of direct weight elicitation [120]. This finding is supported by a separate experiment conducted by Ahn and Park [121].Thus, the ROC method will be applied in the study described in this chapter to elicit weights for the terms in the Rank-Aggregated profile. For a full description of the comparison performed by Barron and Barrett, please see Appendix D.

5.5.2.2 Task 2

The manner in which items are recommended is similar to the process employed by Guo et al. in recommending jobs [30]. This consists of the following steps:

1. Create a candidate set of items to be recommended consisting of the combination of the files in every user's test set.

2. For each user, calculate the mean similarity for each item in the candidate set with the user's training set using the Cosine similarity. For example, suppose a user's training set consists of two items. A candidate item is compared with both items in the training set and has a similarity of 0.2 with the first item and 0.4 with the second item. The test item's mean similarity is 0.3.

3. For each user, rank the items in the candidate set according to mean similarity and select the top 10 items. An item is considered to have been correctly selected for a user if it appears in that user's test set.

Recall that each of the movie and music datasets is split into five different train/test blocks. For each block, recommendation accuracy is computed by averaging the accuracy scores across users. Overall accuracy is computed by averaging accuracy scores across blocks.

5.5.3 Evaluation Method

Each task requires a means for determining whether a structured representation accurately reflects domain category distinctions. If a structured representation accurately reflects domain category distinctions, the items in each category should be more similar to each other than to the items in any other category. For example, in the news dataset, an item in the 'Politics' category should be more similar to the other items in the 'Politics' category than it is the items in the 'Business' category. Recall that the term-list representation is employed in this thesis. Therefore, each item will be represented as a term-list. There are various different methods for determining similarity between term-lists. These can broadly be split into methods that consider weights and methods that do not consider weights. Popular methods that do not consider weights include the Jaccard coefficent [122] and the Dice coefficient [123]. While each of these methods applies a different formula for calculating similarity, each depends solely on *membership* i.e. either a term is present in a list or it is not. However, two different lists may contain the same term, but with different levels of relevance. For example, an article in the 'Business' category may reference a politician who has had an impact on the business community. An article in the 'Politics' category may reference this same politician. However, if domain category distinction have been accurately reflected, the 'Business' article should be more similar to articles in the 'Business' category than to articles in the 'Politics' category, and vice versa. Investigating this property requires a similarity test that accounts for the the different weights that can be assigned to terms within different categories. A much-used method for representing numerical variation between categories is the one-way Analysis of Variance (ANOVA) test [124].

The one-way ANOVA test determines whether a collection of groups differ significantly from each other with respect to a numerical variable. For example, the one-way ANOVA test may be employed to determine whether there is a statistically significant difference between the amount of rainfall in different continents. The one-way ANOVA calculation compares the variance within each group to the variance between groups. The greater the former is than the latter, the more significant is the difference between the groups. In the experiment described in this chapter, domain categories can be considered groups. However, the items in each category are not single numbers, but lists of weighted terms. These items cannot be averaged to create a single number for each category. However, the notion of between-group and within-group variation can be leveraged, using Cosine similarity. The process by which this is acheived is described in the following paragraphs.

For each domain, the following category comparisons are performed:

- Within category - Each individual item in a category is compared with every other item in that category. An average similarity is computed for the item. These averages are averaged to get an average item-to-item similarity for the category.
- Between category - For every category A, each individual item i is compared with every item in every other category. This allows for an average similarity to be computed between each i in A and every other category. These file similarities are averaged to give an average similarity between category A and every other category. Certain items belong to more than one category. In this case, an item is not compared against itself between categories.

Similarity is computed using the Cosine Similarity measure, described in Section 5.5.2.

The 'Within category' and 'Between category' scores for each category are then compared. If, for every category, that category's 'Within category' score is higher than its 'Between category' score, the chosen structured representation has accurately reflected domain category distinctions. Otherwise, it has not.

The quality of generated recommendations is compared using the precision metric, which indicates what percentage of the recommendations were correct. Precision is given by:

$$precision = \frac{correct}{10} \tag{5.3}$$

For example, if a User Model yielded 6 correct recommendations out of a total of 10 its precision score would be 0.6 (6/10). Another commonly used metric in evaluating recommendation quality is *Recall*. Recall is "the fraction of the relevant documents" that are recommended [106]. Recall is not calculated in the experiment described in this chapter because:

1. For task 1, it is not known in advance what the total number of appropriate articles/jobs for the participant will be.
2. Since 10 articles/jobs are recommended in each task, if more than 10 articles/jobs were relevant, it would be impossible to achieve a recall figure of 1.0.

The relationship between similarity scores, diversity and recommendation accuracy

Comparing recommendation precision between user models and recommendation methods will give an indication as to whether certain models or approaches outperform others. However, there may be general patterns with respect to the relationship between properties of generated recommendations and the accuracy of those recommendations. One such pattern is the relationship between similarity scores and recommendation precision i.e. are higher similarity scores associated with recommendation accuracy? However, recall that in Task 1 multiple different user models draw from the same set of candidate news articles/jobs. This allows for the possibility of measuring the relationship between uniqueness (i.e. being recommended by only one model) and recommendation accuracy. This yields the following questions that will be examined:

- Are higher similarity scores correlated with higher recommendation accuracy?

- In task 1, is greater diversity of recommendations correlated with higher recommendation accuracy?

These questions will be investigated using *Logistic Regression*. Logistic Regression measures the relationship between a binary variable of interest (the dependent variable) and multiple numerical and categorical variables (the independent variables) [124]. In this case, the variable of interest is whether a recommended job or news article was correct. The independent variables are the Cosine similarity score of the recommendation and whether or not that recommendation is unique. For each independent variable, Logistic Regression calculates metrics representing the strength and direction of the relationship. In the case of numerical independent variables, direction relates to whether one value of the dependent variable is associated with increases or decreases in the independent variable. In the case of categorical independent variables, direction relates to whether one value of the dependent variable is associated with one value of the independent variable. Logistic Regression calculates the following metrics:

- Coefficient - This value can be positive or negative. The coefficient's sign indicates the direction of the relationship while its value indicates the strength of the relationship.

- P-value - This value represents the probability that the relationship between the independent and dependent variable has occurred by chance. A value of 0.05 (i.e. 5% probability) or lower indicates that the results are statistically significant.

For Task 1, a record is created for each recommended job/news recommendation that represents: (i) The Cosine similarity score of the recommendation (ii) Whether or not the recommendation is unique (iii) Whether or not the recommendation is correct. A Logistic Regression is performed on the collection of records, treating (iii) as the dependent variable. A similar process is employed for Task 2 except that records do not have uniqueness information, since only a single user model is employed. A Logistic Regression is then performed on the record collections.

5.5.3.1 Procedure for Task 1

5 users participated in this evaluation. Participants were identified by means of an email circulated in the research group in which the author works. Similarly to the experiment described in Chapter 4, this is a small number of participants. However, the argument made in Section 4.5 also applies to the experiment described in this chapter. As in the experiment described in Chapter 4, each user in the experiment described in this chapter is a hyperactive Twitter user. This large volume of Twitter data is particularly necessary given the fact that various different user models are being investigated in this experiment.

Each participant was assigned to either the news recommendation task or job recommendation task at random.

After the participant's information has been analysed 10 recommendations are generated for each User Model. The participant is then sent a zip folder containing six subfolders (one for each User Model). Contained in each of these folders are 10 articles/jobs and a separate file in which the participant can indicate whether the article/job is appropriate for them. For the news recommendation task, the participant is asked to indicate whether the article is of interest to them or not. For the job recommendation task, the participant is asked to indicate whether the job is relevant or not. 'Relevant' means that the participant has the expertise required to do the job.

Instructions given to participants

The structured representations of job descriptions in the experiment described in this chapter represent subjects that appear in that job description. However, a person's suitability for a job can depend on factors such as level of experience or having a particular qualification. The method of analysis employed in the experiment described in this chapter does not attempt to identify such information. This could potentially lead to 'false negatives', where for example a participant rejects a job for which they have the required expertise, but for which they do not have the required level of experience. In order to prevent this from occurring, participants were instructed to ignore the level of qualification required for a particular job and focus solely on the core competencies. For example, suppose a job required a participant to be able to program in Java and to have an Oracle certification. If the participant could program in Java but didn't have an Oracle certification, they were instructed to indicate that the job was relevant. Similarly, if a participant had a basic competency required for a job e.g. 'good with Excel' but did not have a core competency e.g. their field of expertise is Computer Science but the job is in the field of Psychology, they were instructed to indicate that the job was not relevant for them.

5.5.4 Results

5.5.4.1 Within-category and Between-category comparisons

News

The news category similarity scores can be seen in Table 5.4. It can be seen from this table that each category's 'Within category' similarity is higher than its 'Between category' similarity with any other category.

	business	entertainment	politics	sport	tech
business	**0.020830**	0.004223	0.013209	0.004383	0.008863
entertainment	0.004223	**0.022675**	0.007363	0.002445	0.007179
politics	0.013209	0.007363	**0.076251**	0.004271	0.008093
sport	0.004383	0.002445	0.004271	**0.017945**	0.004643
tech	0.008863	0.007179	0.008093	0.004643	**0.023996**

Table 5.4: News dataset category similarity scores.

Vacancies

A total of 12 vacancy categories were found to have a greater between-category similarity than within-category similarity. This was most pronounced for the 'Department of Engineering Science', 'Department of Politics and International Relations', 'Humanities Division' and 'University of Oxford' with 27 categories each. Table 5.5 shows a sample of the vacancy categories with higher between-category similarity than within-category similarity.

Category	**Number of categories where between-category similarity is higher**	**Examples**
Department of Engineering Science	27	Bodleian Libraries, Centre for Health Service Economics and Organisation
Mathematical Physical and Life Sciences	8	Personnel and Related Services, Radcliffe School of Medicine
Personnel and Related Services	6	Department of Paediatrics, Mathematical Institute
Social Sciences Division	4	Mathematical Institute, Radcliffe School of Medicine
Mathematical Institute	1	Department of Chemistry

Table 5.5: Vacancy categories with higher 'Between category' similarity than 'Within category' similarity.

Movies

When using AlchemyAPI on the movie dataset, a total of 13 genres were found to have a greater between-category similarity than within-category similarity. This was most pronounced for the Adventure genre, with 13 genres. Table 5.6 shows a sample of the movie genres with higher between-category similarity than within-category similarity when using AlchemyAPI.

Genre	**Number of categories where between-category similarity is higher**	**Examples**
Action	4	Film Noir, Mystery
Adventure	13	Crime, Sci-Fi
Comedy	4	Romance, Western
Horror	2	Film Noir, Western
Thriller	3	Mystery, Western

Table 5.6: Movie genres with higher 'Between category' similarity than 'Within category' similarity, using AlchemyAPI.

When using Babelfy on the movie dataset, a total of 8 genres were found to have a greater between-category similarity than within-category similarity. The genre for which this was most pronounced was Fantasy, with 8 genres. Table 5.7 shows a sample of the movie genres with higher between-category similarity than within-category similarity when using Babelfy.

Genre	**Number of categories where between-category similarity is higher**	**Examples**
Animation	2	Children, Fantasy
Drama	2	Mystery, Romance
Fantasy	8	Sci-Fi, Western
Horror	1	Mystery
Thriller	3	Crime, Film Noir

Table 5.7: Movie genres with higher 'Between category' similarity than 'Within category' similarity, using Babelfy.

Music category similarity

When using AlchemyAPI on the music dataset, a total of 23 genres were found to have a greater between-category similarity than within-category similarity. This was most pronounced for the Karaoke genre, with 31. Table 5.8 shows a sample of the music genres with higher between-category similarity than within-category similarity when using AlchemyAPI.

Genre	**Number of categories where between-category similarity is higher**	**Examples**
Alternative	5	Classical, Indie
Karaoke	31	Alternative, Metal
Reggae	4	Hip Hop, Rap
Rock	4	Indie, Karaoke
Vocal	9	Inspirational, Reggae

Table 5.8: Music genres with higher 'Between category' similarity than 'Within category' similarity, using AlchemyAPI.

When using Babelfy on the music dataset, a total of 25 genres were found to have a greater between-category similarity than within-category similarity. The genre for which this was most pronounced was Comedy, with 32 genres. Table 5.9 shows a sample of the music genres with higher between-category similarity than within-category similarity when using Babelfy.

Genre	**Number of categories where between-category similarity is higher**	**Examples**
Blues	14	Indie, Reggae
Comedy	32	Pop, Soul
Industrial	22	Country, Reggae
Jazz	25	Electronic, R&B
Soul	6	Karaoke, New Age

Table 5.9: Music genres with higher 'Between category' similarity than 'Within category' similarity, using Babelfy.

The full list of genres which were found to have a greater between-category similarity than within-category similarity, as well as the similarity scores for each domain and analysis tool, can be found here: `https://github.com/amcgover/PhD-Data`.

5.5.4.2 Recommendation precision

Task 1

The precision scores for the news and jobs recommendation tasks are given in Table 5.10 and 5.11 respectively. It can be seen that there is no clear performance benefit for any of the User Models employed. It can also be seen that the precision scores for the news recommendation task are significantly higher than those for the job recommendation task.

	User Model 1	User Model 2	User Model 3	User Model 4	User Model 5	User Model 6
Participant 1	0.6	0.8	0.8	0.75	0.6	0.8
Participant 4	0.9	0.9	0.9	1.0	0.7	0.8

Table 5.10: News recommendation precision percentages.

	User Model 1	User Model 2	User Model 3	User Model 4	User Model 5	User Model 6
Participant 2	0.3	0.2	0.2	0.25	0.1	0.2
Participant 3	0.2	0.1	0.2	0.0	0.1	0.0
Participant 5	0.3	0.2	0.3	0.4	0.0	0.4

Table 5.11: Job recommendation precision percentages.

Task 2

The recommendation precision achieved on the music and movie datasets using AlchemyAPI can be seen in Table 5.12.

Dataset	**Batch 1**	**Batch 2**	**Batch 3**	**Batch 4**	**Batch 5**	**Average**
Music	0.035400	0.035700	0.036800	0.035000	0.034800	0.03554
Movies	0.009586	0.010296	0.013018	0.010296	0.010296	0.01070

Table 5.12: Movie and Music Recommendation precision, using AlchemyAPI.

The recommendation precision achieved on the music and movie datasets using Babelfy can be seen in Table 5.13. Note that recommendation precision is low for each of these domains.

Dataset	**Batch 1**	**Batch 2**	**Batch 3**	**Batch 4**	**Batch 5**	**Average**
Music	0.012100	0.014500	0.014900	0.012200	0.011400	0.01302
Movies	0.012915	0.013981	0.013863	0.013270	0.015047	0.01382

Table 5.13: Movie and Music Recommendation precision, using Babelfy.

5.5.4.3 The relationship between similarity scores, diversity and recommendation precision

Task 1

Tables 5.14 and 5.15 show the coefficients and p-values for uniqueness and Cosine similarity scores of recommendations and the accuracy of those recommendations in the news and job recommendation tasks, respectively. The p-values in each of these tables indicate that the results are not statistically significant.

Variable	**P-Value**	**Coefficient**
Cosine Similarity	0.914	-0.7740
Unique (True or False)	0.270	-0.7618

Table 5.14: Logistic Regression results for the relationship between similarity scores, diversity and recommendation accuracy for the news reoommendation task.

Variable	**P-Value**	**Coefficient**
Cosine Similarity	0.835	-1.1165
Unique (True or False)	0.053	-1.0921

Table 5.15: Logistic Regression results for the relationship between similarity scores, diversity and recommendation accuracy for the job recommendation task.

Task 2

Tables 5.16 and 5.17 show the coefficients and p-values for the Cosine similarity scores of recommendations and the accuracy of those recommendations for each batch of the music recommendation task for AlchemyAPI and Babelfy, respectively. The p-values in Table 5.16 indicate that the relationship between similarity score and recommendation accuracy is statistically significant for each batch. The coefficient values in Table 5.16 indicate that higher similarity scores are associated with correctness of recommendations. The p-values from Table 5.17 indicate that only the figures for batches 3 and 4 are statistically significant. The coefficients for these values indicate that lower similarity scores are associated with correctness of recommendations.

Batch	**P-Value**	**Coefficient**
1	0.000	26.839
2	0.000	30.2446
3	0.000	30.7061
4	0.000	30.8998
5	0.000	27.7872

Table 5.16: Logistic Regression results for the relationship between similarity scores and recommendation accuracy for the music recommendation task with AlchemyAPI.

Batch	**Cosine Similarity P-Value**	**Cosine Similarity Coefficient**
1	0.147	-6.2812
2	0.154	-5.5379
3	0.012	-9.0435
4	0.036	-9.4246
5	0.246	-5.0523

Table 5.17: Logistic Regression results for the relationship between similarity scores and recommendation accuracy for the music recommendation task with Babelfy.

Tables 5.18 and 5.19 show the coefficients and p-values for the Cosine similarity scores of recommendations and the accuracy of those recommendations for each batch of the movie recommendation task for AlchemyAPI and Babelfy, respectively. The p-values in both tables indicate that the results in these tables are statistically significant. The negative coefficient values indicate that lower similarity scores are associated with correctness of recommendations.

Batch	P-Value	Coefficient
1	0.006	-16.4128
2	0.003	-17.4676
3	0.015	-13.3522
4	0.002	-18.0104
5	0.004	-17.1339

Table 5.18: Logistic Regression results for the relationship between similarity scores and recommendation accuracy for the movie recommendation task with AlchemyAPI.

Batch	Cosine Similarity P-Value	Cosine Similarity Coefficient
1	0.000	-30.5191
2	0.000	-34.9867
3	0.000	-32.7028
4	0.000	-36.1673
5	0.000	-32.8781

Table 5.19: Logistic Regression results for the relationship between similarity scores and recommendation accuracy for the movie recommendation task with BabelNet.

5.6 Discussion

5.6.1 The relationship between the within-category/between-category test and recommendation accuracy

For the vacancy, music and movie datasets, categories were identified with higher between-category similarity than within-category similarity. Note that recommendation accuracy was extremely low for each of these domains and, in the case of the movie and music domains, for both analysis tools. However, no genre in the news dataset had a higher between-category similarity than within-category similarity. These results indicate that there is a direct relationship between a structured representation's ability to accurately reflect domain category distinctions and the accuracy of recommendations generated using that representation. That is, in domains where each category's 'Within category' similarity is greater than its 'Between category' similarity with any other category, it appears that items can be recommended with high precision. In domains where this property does not hold, recommendation precision is severely degraded. This subsection will investigate different ways in which the structured representations employed in this experiment have violated domain category distinctions. For each domain (i.e. news, jobs, music, movies) the discussion will focus on the 10 highest-weighted subjects in each category. The average number of non-unique subjects per category will be calculated in each domain. A subject is non-unique if it appears in at least one other category. This calculation forms the basis of an examination of the extent to which subject-list descriptions are unique for each category. The files containing the results of this calculation for each domain and analysis tool (i.e. AlchemyAPI and Babelfy) can be found here: `https://github.com/amcgover/PhD-Data/tree/master`

News

The average number of non-unique subjects in the top-10 lists of each news category is 0.1, or approximately one subject out of ten. The non-unique subjects were 'Law by issue' (shared by Business and Politics) and 'Written communication' (shared by Business, Politics and Technology). These results show that the Entertainment and Sports news categories have completely unique top-10 subject lists. These results also show that there is a slight similarity between the Business and Politics categories, and a slightly weaker similarity between these two categories and Technology. This suggests that articles in these categories have a certain overlap. However, this overlap is not so great that it affects the overall category distinctions.

Vacancies

The average number of unique subjects in the top-10 lists of each vacancy category was 0.56, or approximately six subjects out of ten. Table 5.20 shows examples of non-unique subjects. Table 5.20 also shows the number of vacancy categories in which the subjects appear, as well as example categories.

Subject	**Total number of departments in which subject appears in Top 10**	**Examples**
Academia	10	Department of Chemistry, Mathematical Physical and Life Sciences (MPLS)
Clinical research	7	Medical Sciences Division, Nuffield Department of Population Health, Personnel and Related Services
Knowledge	7	Humanities Division, Mathematical Physical and Life Sciences (MPLS), Nuffield Department of Clinical Medicine
Culture in Oxford	6	Department of Paediatrics, Department of Physics, Humanities Division

Table 5.20: Non-unique subjects in the Vacancy domain.

Note the generic nature of the subjects in Table 5.20. These subjects are a combination of subjects relating to academia in general and the University of Oxford in general.

Movies

For the AlchemyAPI tool, the average number of non-unique subjects in each movie genre is 0.72, or approximately seven subjects out of ten. Table 5.21 shows examples of non-unique subjects in the Movie domain for AlchemyAPI. Table 5.21 also shows the number of genres in which the subjects appear, as well as example genres. For Babelfy, the average number of non-unique subjects was 0.78, or approximately eight subjects out of ten. Table 5.22 is an analogue of Table 5.21.

Subject	**Total number of genres in which subject appears in Top 10**	**Examples**
English-language films	18	Action, Comedy, Horror
Family	13	Drama, Romance, Western
Love	7	Animation, Children, Mystery
Interpersonal relationship	5	Children, Fantasy, Musical

Table 5.21: Non-unique subjects in the Movie domain for the AlchemyAPI analysis tool.

Subject	**Total number of genres in which subject appears in Top 10**	**Examples**
Film	15	Documentary, Film Noir, Western
Human	14	Horror, Sci Fi, War
Woman	9	Drama, Romance, Western
Family	6	Children, Horror, Musical

Table 5.22: Non-unique subjects in the Movie domain for the Babelfy analysis tool.

Note the generic nature of the subjects for each of the analysis tools. That the 'English-language films' and 'Film' subjects are generic is trivially true. However, subjects such as 'Family' and 'Interpersonal relationship' convey very little genre-specific information. Instead, they appear to signify more general information about the nature of movies as stories. That is, many movies in different genres appear to contain themes relating to these subjects.

Music

For the AlchemyAPI tool, the average number of non-unique subjects in each music genre is 0.76, or approximately eight subjects out of ten. The total number of non-unique subjects is 34. For Babelfy, the average number of non-unique subjects was 0.94, or approximately nine subjects out of ten. The number of non-unique subjects between genre top-10 lists is 24. Tables 5.23 and 5.24 are analogues of Tables 5.21 and 5.21, respectively.

Subject	**Total number of genres in which subject appears in Top 10**	**Examples**
Song	33	Alternative, Country, Metal
Music	31	Blues, Easy Listening, Folk
Singing	27	Funk, Indie, Rap
Gramophone record	27	Electronic, Instrumental, Jazz

Table 5.23: Non-unique subjects in the Music domain for the AlchemyAPI analysis tool.

Subject	**Total number of genres in which subject appears in Top 10**	**Examples**
Rock Music	26	Classical,Industrial, Reggae
Song	19	Dance, Rock, World
Album	7	Hip Hop, Jazz, Vocal
Single	5	Alternative, Karaoke, Soul

Table 5.24: Non-unique subjects in the Music domain for the Babelfy analysis tool.

Note again the generic nature of the subjects. Subjects such as 'Song', 'Album' and 'Singing' relate to the Music domain in general, and do not reveal genre-specific information.

The preceding discussion clearly demonstrates an important issue with employing structured representation in particular personalisation scenarios. In certain cases, a structured representation captures generic subjects. This causes spurious similarities to be represented between categories. The experiment described in this chapter demonstrates how this issue affects recommendation. Note that this property is not revealed by studies such as Abel et al. [7], Guo et al. [30] and Li et al. [62]. Such approaches compare the results of employing different structured representations in recommendation tasks. However, these approaches do not consider the extent to which these structured representations accurately reflect domain category distinctions. The results of the experiment described in this chapter - as well as preceding discussion - demonstrates that this consideration can have a significant effect on recommendation accuracy.

It should be noted that there are naturally other domains for recommendation[16]. Thus, it cannot be stated to a certainty that these findings will occur in every recommendation domain. However, these findings were demonstrated across four distinct domains, using multiple different analysis methods. These findings are thus indicative, but not definitive.

5.6.2 The relationship between similarity scores, diversity and recommendation precision

Note the inconsistency in the results given in Tables 5.14 to 5.19. Only the music and movie recommendation tasks yield p-values indicating statistical significance. In the movie domain, the coefficient directions indicate that lower similarity scores are associated with correctness of recommendations. This effect can also be seen in the music domain for the Babelfy analysis tool. This directly contradicts the interpretation of Cosine similarity as applied in personalisation approaches such as those referenced in Section 5.5.2 [7], [104], [117], [118] i.e. that higher Cosine similarity scores are associated with greater accuracy.

These results indicate that the gap in recommendation accuracy between domains in the experiment described in this chapter is not attributable to uniqueness or levels of Cosine similarity.

5.6.3 Points to note

For no participant did the Knowledge-specific User Model produce the full ten recommendations i.e. the number of jobs/news articles for which a non-zero Cosine Similarity score was obtained was less than ten for each participant. This explains the 1.0 precision figure for Participant 4, whose Knowledge-specific User Model only produced one recommendation. This suggests that there are certain sources of user information that have a particularly narrow scope in which that information can be applied.

For no participant did the Schulze algorithm find an outright winner. This means that each item in the Schulze-aggregated User Model had the same weight. This also means that for each participant the Interest-specific User Model and Knowledge-specific User Model were generated using the first-ranked interests in the user's tweets and publications, respectively. This suggests that when two user profiles are being aggregated, the Schulze Method offers no benefits in aggregation.

[16] Park et al. identify eight recommendation domains [125]. These are: books, documents, images, movie, music, shopping, TV programs, and others.

Some of the categories in the Oxford jobs dataset were highly specific e.g. Bodleian Libraries, Nuffield Department of Clinical Medicine. It could be argued that it is unreasonable to expect a structured representation to reflect category distinctions that are so specific, and therefore that it might be expected that the between-category/within-category test would fail. However, if failure of the test were only due to over-specificity of categories, one would expect that job recommendations would still be broadly accurate. For example, while a structured representation may not be expected to recommend a specific Oxford department for a Computer Scientist, it should still recommend jobs in containing a computer science component or another component such as mathematics. This was not the case for the vacancy recommendations provided. Participants received jobs in areas such as Journalism, Climate Change Research and Population Health.

5.7 Conclusion

This chapter described an experiment investigating the effect of content domain on recommendation accuracy. A test was proposed for determining whether a structured representation accurately reflects domain category distinctions. This test compares average similarity between items in order to determine whether items in the same category are more similar to each other than they are to items in different categories. Multiple recommendation tasks were performed spanning a variety of domains, user models and recommendation methods. The results of these tasks indicate that this test can identify when a structured representation will be able to facilitate effective recommendation.

It should be noted that the between-category/within-category test applied in the experiment described in this chapter depends on the representation format and similarity metric employed. The term-list format and Cosine similarity metric are widely used in personalisation approaches, giving the results of this experiment substantial generalisability. However, other representation formats and similarity metrics exist. One should therefore not assume that the test will perform similarly with different representation formats and similarity metrics.

Chapter 6

Conclusion

6.1 Summary of thesis contributions

Recall that the Research Question of this thesis is as follows:

How does the choice of structured representation[1] *affect the personalisation that can be provided using that representation?*

There are three subsections in this section: one for each of the characteristics described on page 6 of this thesis. Each subsection describes:

- The experiment that was conducted to investigate the characteristic
- The findings of that experiment
- The contributions of those findings

Characteristic 1 - The information the structured representation contains

This characteristic was investigated through an experiment examining how the inability of content-based representations such as Named Entities and Bag-Of-Words to capture context limits their ability to represent different forms of user expression. For each form of user expression examined, this experiment clearly identified the inability of these representations to fully describe that form of expression. This is in contrast to approaches such as Hecht et al. that employ the Named Entity representation to provide personalisation without

[1] Where the choice of structured representation is characterised according to the dimensions given on page 6.

investigating how their choice of representation affected personalisation performance [6]. This is also in contrast to approaches such as Michelson and Macskassy [21] and Kapanipathi et al. [22] that argue that the Named Entity representation's limitations can be overcome by categorising Named Entities. Such approaches do not provide insight as to whether or not this categorisation approach addresses the fundamental limitations of the Named Entity representation.

Characteristic 2 - The external linguistic resource employed to generate the structured representation

This characteristic was investigated through an experiment examining how the social relevance of an external linguistic resource affects that resource's ability to provide the context that is absent in content-based representations such as Named Entities and Bag-Of-Words. This experiment took the form of a comparison of a purely lexical resource (WordNet) and a general knowledge base (DBpedia) in representing user interests and knowledge. This experiment produced indicative findings that the greater social relevance of DBpedia makes it more suitable than WordNet for representing user interests and knowledge. This is in contrast to approaches such as Orlandi et al. that leverage an external linguistic resource to perform personalisation without examining how the hierarchy provided by their resource affected personalisation performance. Orlandi et al. approximate such an investigation by asking participants to enumerate a list of subjects that are relevant for them [29]. These authors then calculate the proportion of this list that is covered by their approach. However, Orlandi et al. admit that this approach cannot give a reliable answer. This is especially the case since only one external linguistic resource is employed, providing no basis for comparison.

Characteristic 3 - The domain of the unstructured content being analysed

This characteristic was investigated through an experiment examining how a structured representation's ability to accurately reflect domain category distinctions affects that representation's ability to provide effective recommendation. This experiment identified a straightforward test that can be performed to determine whether a structured representation accurately reflects domain category distinctions. If a structured representation fails this test, recommendation precision will be severely degraded. This experiment also identified reasons why a structured representation may fail this test. These findings were demonstrated across multiple domains, user models and recommendation methods. This is in contrast to approaches such as Guo et al. [30] and Li et al. [62] that investigate the efficacy of different structured representations in personalisation tasks. Such approaches only examine information such as recommendation accuracy. If a structured representation A yields better personalisation performance than a structured representation B, A must have captured some important information that B did not. Approaches such as Guo et al. and Li et al. do not investigate what this information is, and why one structured representation captures it when another does not.

The experiments described in this thesis thus closely examine and elucidate the significance of these three characteristics for personalisation approaches. Taken toegther, the findings of these experiments provide indicative guidelines as to how to determine whether a content-based representation is suitable for the personalisation task at hand. If such a representation is not suitable, these findings provide guidelines as to how to choose an external linguistic resource that will address the deficiency of that representation. These findings clearly illustrate the effects of the personalisation task at hand on the choice of structured representations, as well as how these effects should be addressed by personalisation approaches.

6.2 Future work

This section discusses areas for exploration, based on the results of the experiments described in this thesis.

Addressing the within-category/between-category issue

The experiment described in Chapter 5 described a method for determining whether structured representations reflect domain category distinctions. This experiment demonstrated that when structured representations do not reflect domain category distinctions, recommendation precision will be severely degraded. A natural area for future work concerns how to proceed if structured descriptions do not reflect domain category distinctions. One way to address this issue would be to remove subjects so that each category's within-category similarity is greater than its between-category similarity with any other category. Two manners in which this could be performed are:

1. Over the whole dataset - Subjects that appear in a majority of categories could be removed. For example, if a subject appears in 60% of categories in a domain, it could be considered generic and non-discriminative, and could be removed.

2. Pairwise - Each category could be compared with the categories with which it has higher between-category similarity than within-category similarity. Subjects that the categories share would be identified. For each subject, the task would be to determine for which

of the two categories the subject is most relevant. This could be determined by: (i) Comparing the ratio of files in which the subject appears for each category. For example, if the term appears in 50% of the files in category A and 60% of the files in category B, the subject is more relevant for B than A. (ii) Comparing the average weight of the subject for each category. For example, if the subject's average weight is *0.7* in category A and *0.3* in category B, the subject is more relevant for A than B.

Note that it is likely that the above process would need to be tuned according to the domain in which it was being applied. For example, when applying approach 1 'majority' could be defined as 60%, 70% etc.

While the above approaches describe ways for adapting a particular structured representation, there is another potential approach for addressing the within-category/between-category issue. Redefining domain categories is another way in which spurious similarities could be resolved. For example, if there is a large amount of shared information between the structured representations for two categories, these categories could be merged. Recall however from the discussion in Section 5.6 that the spurious similarity in the case of the music and movie datasets arose as a result of overly generic subjects being identified. Thus, while it may be advisable to redefine domain categories in certain situations, it is not advisable for the situation described in this thesis.

Refining the within-category/between-category test

Recall that the within-category/between-category test described in Chapter 5 makes reference to the ANOVA test. Associated with ANOVA are a p-value and an effect size. The former has the same interpretation as the p-value described with respect to Logistic Regression in Section 5.5.3. Effect size takes values from 0 to 1, and refers to the extent of the difference between groups. A value of 1 indicates that the groups have no similarity whatsoever while a value of 0 indicates that the groups are virtually indistinguishable from one another. The test described in Chapter 5 could be expanded to experiments to investigate whether p-values and/or effect sizes could be reliably calculated. In order to be robust, these experiments would need to incorporate multiple representation formats and similarity metrics.

6.3 Concluding Remarks

This thesis described a series of experiments investigating the application of different structured representations of unstructured content for the purpose of personalisation. Specific structured representations and different external linguistic resources were examined. The influence of content domain on the efficacy of structured representations in personalisation tasks was also examined.

Appendix A

The WordNet Domains hierarchy with corresponding Dewey Decimal Classification Codes

The following domains are not mapped to a Dewey Decimal Classification (DDC) code:

1. Factotum
2. Number
3. Color
4. Time Period
5. Person
6. Quality
7. Metrology
8. Psychological Features

Each of the domains 2-8 is a sub-domain of the ‘Factotum’ domain, which is applied to synsets for which no specific domain can be found. Thus, these domains have no corresponding DDC code.

	Basic Domains			DDC
Humanities				
	History			[920:990]
		Archaeology		930.1
		Heraldry		[929.6,929.7]
	Linguistics			410
		Grammar		415
	Literature			[800,400]
		Philology		400
	Philosophy			[100-(130,150,176)]
	Psychology			150
		Psychoanalysis		150.195
	Art			[700-(710,720,745.5,790-(791.43,792,793.3))]
		Graphic_Arts		760
			Philately	769.56
		Dance		[792.8,793.3]
		Drawing		[740-745.5]
		Painting		750

		Music		780
		Photography		770
		Plastic_Arts		730
			Jewellery	739.27
			Numismatics	737
			Sculpture	[731:735]
		Theatre		[792-792.8]
		Cinema		791.43
	Paranormal			130
		Occultism		[133-133.5]
		Astrology		133.5
	Religion			200
		Theology		[220,230,240,260]
		Roman_Catholic		282
		Mythology		291.13
Free_Time				[790-(791.43,792,793.3)]
	Radio-Tv			[791.44,791.45]
	Play			[793.4:795-794.6]
		Betting		[795.1:795.3]
		Card		795.4

		Chess		794.1
	Sport			[794.6,796:799]
		Badminton		796.345
		Baseball		796.357
		Basketball		796.323
		Cricket		796.358
		Football		796.332
		Golf		796.352
		Rugby		796.333
		Soccer		796.334
		Table_Tennis		796.346
		Tennis		796.342
		Volleyball		796.325
		Cycling		796.6
		Skating		796.21
		Skiing		796.93
		Hockey		796.962
		Mountaineering		796.522
		Rowing		797.122
		Swimming		797.21
		Sub		797.23
		Diving		797.2
		Racing		796.72
		Athletics		796
		Wrestling		796.812
		Boxing		796.83
		Fencing		796.86
		Archery		799.32
		Fishing		799.1

		Hunting		799.2
		Bowling		794.6
Applied_Science				600
	Agriculture			[630,338.1]
		Animal_Husbandry		[636,638,639]
			Veterinary	636.089
	Food			[613.2,613.3,641,642]
		Gastronomy		[642,641.5:641.8-641,62]
	Home			[640-(641,642,645)]
	Architecture			[645,690,710,720]
		Town_Planning		710
		Buildings		690
		Furniture		645
	Computer_Science			[004:006]
	Engineering			620

		Mechanics		620.1
		Astronautics		629.4
		Electrotechnology		621.3
		Hydraulics		627
	Telecommunication			[383,384]
		Post		383
		Telegraphy		384.1
		Telephony		384.6
	Medicine			[610-(611,612,613)]
		Dentistry		617.6
		Pharmacy		615
		Psychiatry		616.89
		Radiology		616.075
		Surgery		[617-617.6]
Pure_Science				500
	Astronomy			520
	Biology			[570-577,611,612-612.6]
		Biochemistry		572
		Anatomy		611
		Physiology		[571,612,-612.6]
		Genetics		576

	Animals			590
		Entomology		595.7
	Plants			580
	Environment			577
	Chemistry			540
	Earth			[550,560,910-(910.4,910.202)]
		Geology		[551-(551.5,551.6,551.46)]
		Meteorology		[551.5,551.6]
		Oceanography		551.46
		Paleontology		560
		Geography		[910-(910.4,910.202)]
			Topography	912
	Mathematics			510
		Geometry		516
		Statistics		519.5

	Physics			530
		Acoustics		534
		Atomic_Physic		539.7
		Electricity		537
			Electronics	537.5
		Gas		533
		Optics		535
Social_Science				[300.1:300.9]
	Anthropology			[301:307,395,398]
		Ethnology		305.8
			Folklore	398
	Health			[613-(613.2,613.3,613.8,613.9)]
		Body_Care		613.4
	Military			[355:359]
	Pedagogy			370
		School		[371:373]
		University		378
	Publishing			070

	Sociology			[301:319-(305.8,306.7),360-(363.4,368)]
	Artisanship			[745.5,338.642]
	Commerce			[381,382]
	Industry			[338-(338.1,338.642),660,670,680]
	Transport			[385:389]
		Aviation		[387.7,387.8]
		Vehicles		[388-388.1]
		Nautical		[386,387.1:387.5]
		Railway		385
	Economy			[330-338,368,650]
		Enterprise		650
			Book_Keeping	657
		Finance		[332,336.3]
			Banking	[332.1:332.3]
			Money	332.4
			Exchange	332.6
		Insurance		368
		Tax		[336.1,336.2]

	Administration			[351:354]
	Law			340
	Politics			320
		Diplomacy		327.2
	Tourism			[910.202,910.4]
	Fashion			[390-(392.6,395,398),687]
	Sexuality			[155.3,176,306.7,363.4,392.6,612.6,613.96,]

Appendix B

Word Sense Disambiguation

There are often multiple possible interpretations for a word in text. For example, the word 'bank' has entirely different interpretations when it appears to the right of the word 'river' than when it appears to the right of the word 'merchant'. The NLP technique Word Sense Disambiguation (WSD) addresses this issue. Lee and Ng define Word Sense Disambiguation as follows: "Given an occurrence of a word w in a natural language text, the task of word sense disambiguation (WSD) is to determine the correct sense of w in that context" [126]. WSD has been applied in tasks as diverse as recommending talks at conferences [40] and implementing a Computer Aided Software Engineering System [100]. WSD will be applied in the case study described in Chapter 4, which compares the DBpedia and WordNet resources. The following subsections describe the WSD approaches that will be employed in this study with respect to WordNet and DBpedia.

B.1 WordNet

Magnini et al. [92] perform WSD using WordNet Domains. WordNet Domains[1] is a resource that assigns domain labels to WordNet synsets in accordance with the Dewey Decimal Classification (DDC) [24], [127]. Each WordNet Domains label corresponds to one or more DDC codes. For example, the WordNet Domain 'Law' corresponds to the DDC code 340[2]. Magnini et al. first determine the mean and standard deviation for every WordNet domain in the LOB corpus[3]. They then use this information to determine the relevant domains for the portion of text in which the ambiguous term appears. This list of domains is referred to as the *text vector*. What follows is an example the authors describe for creating a text vector. The contents of 'Table 1' can be found in Table B.1.

[1] http://wndomains.fbk.eu/index.html

[2] For the complete WordNet Domains hierarchy as well as corresponding DDC codes, please see Appendix A.

[3] http://khnt.hit.uib.no/icame/manuals/lob/

Sense	Synset and Gloss	Domains
1	depository financial institution, bank, banking concern, banking company (a financial institution ...)	Economy
2	bank (sloping land ...)	Geography, Geology
3	bank (a supply or stock held in reserve ...)	Economy
4	bank, bank building (a building ...)	Architecture, Economy
5	bank (an arrangement of similar objects ...)	Factotum
6	savings bank, coin bank, money box, bank (a container ...)	Economy
7	bank (a long ridge or pile ...)	Geography, Geology
8	bank (the funds held by a gambling house ...)	Economy, Play
9	bank, cant, camber (a slope in the turn of a road ...)	Architecture
10	bank (a flight maneuver ...)	Transport

Table B.1: Table 1 from Magnini et al.'s WSD paper.

"For example, suppose we want to evaluate the relevance of Economy in the sentence 'Today I draw money from my bank'. The algorithm will collect all the domains for each sense of each word. The noun 'bank' has five occurrences of Economy in its synsets out of a total of 10 senses (see Table 1), the noun 'money' has three occurrences out a total of three, and the verb 'draw' has one occurrence out a total of 33. Then the total frequency of Economy is 1.53. Suppose that the mean frequency of Economy for texts of this length in the LOB corpus is 0.2 and that the standard deviation is 0.1. These values represent the frequency distribution of Economy in unrelated texts. As a consequence, Economy will not be relevant in texts in which its frequency is in the range [0, 0.4], while it will be considered relevant in texts with significantly higher frequency, as it is the case of our example"

Once the text vector has been constructed for the context in which the ambiguous term appears, *sense vectors* are constructed for the ambiguous term. A sense vector consists of the domain labels of that particular sense, and is of length proportional to the number of times that sense occurs in the SemCor corpus[4]. SemCor is a collection of articles from the Brown Corpus in which each word has been manually annotated with its correct sense. For example, Sense 1 of 'bank' in Table B.1 appears twenty times in SemCor, so its sense vector is consists of the value *20* for the Economy domain and 0 for every other domain.

The dot product between each sense vector of the ambiguous term and the domain vector of the text is then computed. This provides a ranking of the senses according to their domain similarity to the context. This ranking is then used to choose the correct sense.

However, WordNet Domains labels are mapped to WordNet 1.6. WordNet has since been updated with new synsets. Therefore, in the analysis described in this thesis, the Extended WordNet Domain labels created by González-Agirre et al. [84] are used. The Extended WordNet Domains are mapped to WordNet 3.0 synsets. González-Agirre et al. report better WSD performance with Extended WordNet Domains than with the original WordNet Domains.

B.2 DBpedia

Mihalcea and Csomai describe an approach for identifying the relevant Wikipedia articles for a piece of text [128]. In this approach, senses are represented by Wikipedia article titles. For a given text, the authors first extract *keyphrases* in the text. A keyphrase is a single- or multi-word phrase that "expresses the primary topics and themes" of a text [87]. A list of keyphrases for a text constitutes a summary of the content of that text. With respect to Wikipedia articles, Mihalcea and Csomai define keyphrases as links to other Wikipedia articles. In order to identify keyphrases, Mihalcea and Csomai construct a controlled vocabulary consisting of Wikipedia article titles. However, a Wikipedia article can be linked using text other than its title. For example, the text 'java' can link to the Wikipedia article 'Java (programming language)'. The text linking to an article is called that article's 'surface form'. Mihalcea and Csomai extend their vocabulary with the surface forms from all Wikipedia articles, discarding the surface forms that appear less than five times in the Wikipedia corpus. This controlled vocabulary is then used to identify keyphrases in text.

However, a text may contain many phrases that are in the controlled vocabulary. If the ratio of the number of keyphrases to the total number of words in the text is too high, the keyphrase

[4] `https://www.gabormelli.com/RKB/SemCor_Corpus`

list ceases to be a summary. Thus, an appropriate ratio of keyphrases to words in the text must be determined. In order to determine this ratio, Mihalcea and Csomai analyse the entire Wikipedia corpus. The authors find that the average ratio of keyphrases to words per Wikipedia article is 6%. Identifying keyphrases thus consists of three steps:

1. Identify all phrases in the text that are in the controlled vocabulary. Each of these keyphrases is a *candidate keyphrase*. Example candidate keyphrases are 'Java', 'World Cup' etc.

2. Assign a numeric value to each candidate keyphrase that reflects the likelihood of that candidate keyphrase being selected as a keyphrase for the text. Then rank the candidate keyphrases.

3. Select the top 6% ranked candidate keyphrases as the keyphrases for the text.

Mihalcea and Csomai evaluate three different methods for assigning the numeric weights for step 2. One of these is tf-idf, which is described in Section 2.3.3. The other two ranking methods are described in the following subsections.

The x^2 independence test

Also known as the chi-square test, this measures whether a candidate keyphrase *p* occurs more frequently in a particular document *d* than it does in the corpus in general. Mihalcea and Csomai use the collection of Wikipedia articles as a corpus. The first step in this process is the creation of a table describing phrase occurrence frequencies in *d* and the entire corpus. This table can be seen in Table B.2.

freq(p, d)	*freq(CKP_less_p, d)*
freq(p, Corp_less_d)	*freq(CKP_less_p, Corp_less_d)*

Table B.2: Information required for x^2 independence testing.

freq(p, d) refers to the number of times the phrase *p* appears in *d*. *freq(p, Corp_less_d)* refers to the total number of times *p* appears in all documents in the corpus, excluding *d*. *freq(CKP_less_p, d)* refers to the sum of the frequencies of every candidate keyphrase, excluding *p*, in *d*. *freq(CKP_less_p, Corp_less_d)* refers to the sum of the frequency counts for each candidate keyphrase, excluding *p*, in every document in the corpus, excluding *d*.

5	43	48
71	235	306
76	278	354

Table B.3: Sample values for the cells in Table B.2

Example values can be seen in Table B.3. The third cell in each row contains the total for that row. The third cell in each column contains the total for that column. This table conveys the following information:

- p occurs 5 times in d.
- The total frequency of identified keyphrases, excluding p, in d is 43.
- p occurs 71 times in the whole corpus, excluding d.
- The total frequency of d's candidate keyphrases ,excluding p, in the whole corpus, excluding d, is 235.

For each cell in this table an *expected value* is computed. This value represents the value the cell would have if p does not occur more in d than it does in the corpus on average. The formula for calculating the expected value for a cell is as follows:

$$\text{expected value} = \frac{\text{row total} \cdot \text{column total}}{\text{n}} \tag{B.1}$$

where n is the overall total i.e. the value in the third column of the third row. So, for the first cell (i.e. row one, column one) of Table B.3, the expected cell count is given by:

$$\text{expected value} = \frac{48 \cdot 76}{354} = 10.31 \tag{B.2}$$

Once the expected counts have been calculated for each cell, the x^2 statistic is calculated as the sum of the squares of the differences between each cell's observed value and its expected value. So, for Table B.3, the x^2 statistic is 4.02. The higher a candidate keyphrase's x^2 statistic, the greater its likelihood of being selected as a keyphrase.

Keyphraseness

This measure is given by the following formula:

$$score = \frac{\text{apps\_as\_link}}{\text{tot\_apps}} \tag{B.3}$$

where *apps_as_link* refers to the number of documents in which the candidate keyphrase appears as a link and *tot_apps* refers to the total number of documents in which the candidate

Method	Precision	Recall	F-measure
tf-idf	41.91	43.73	42.82
x^2 statistic	41.44	43.17	42.30
keyphraseness	**53.37**	**55.90**	**54.63**

Table B.4: Mihalcea and Csomai's results from comparing the tf-idf, x^2 and keyphraseness ranking methods.

keyphrase appears. For example, the word 'the' will receive a low keyphraseness score, as it appears in a large number of the articles in Wikipedia, and in the large majority of these it does not appear as a link.

The authors compare the tf-idf, x^2 and keyphraseness ranking measures by evaluating how effective each is in identifying keyphrases from a manually annotated set of 100 Wikipedia articles. The authors use the following metrics in their comparison:

- Precision - This is the ratio of the number of keyphrases a method correctly identified to the total number of keyphrases the method identified. For example, if a method identified 2000 keyphrases of which 1000 were correct its precision would be 50%.
- Recall - This is the ratio of the number of keyphrases that a method correctly identified to the total number of keyphrases that were manually annotated. For example, if 1500 keyphrases had been manually annotated and a method identified 1000 of these its recall would be 66% (with rounding).
- F-measure - Also known as F1, this is a weighted average of precision and recall. It is given by the following formula:

$$\text{F-measure} = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{B.4}$$

  For example, with the example precision and recall figures given in the previous two bullet points, the F-measure would be approximately 57%.

The results of comparison of the tf-idf, x^2 and keyphraseness ranking methods are given in Table B.4. These results show that the keyphraseness ranking method outperforms the other two ranking methods.

Mihalcea and Csomai evaluate two different approaches for linking keyphrases with the correct Wikipedia article. Both involve comparing the context in which the keyphrase appears with

Sense Number	Definition
1	A solid or hollow object which tapers from a circular or roughly circular base to a point
2	A conical mountain, especially one of volcanic origin
3	A plastic cone-shaped object that is used to separate off or close sections of a road

Table B.5: Possible definitions for the word ‘cone’.

the articles to which the keyphrase can possibly refer. Mihalcea and Csomai define ‘context’ as the paragraph in which the ambiguous phrase appears. Both methods are described in the following subsections.

Knowledge-based

This algorithm is a simplified version of the algorithm defined by Lesk [129]. In the simplified Lesk algorithm, an ambiguous phrase is disambiguated by counting the number of words shared by the context in which the phrase appears and the phrase's possible dictionary definitions. The dictionary definition that shares the most words with the context is chosen as the correct interpretation of the phrase. Before counting the number of overlaps, stop words and punctuation are removed. For example, consider the following sentence: *‘I found a cone on the road outside my house.’*. Suppose the phrase to be disambiguated is ‘cone’. Table B.5 shows some possible definitions for ‘cone’[5]. Sense 3 would be chosen because of the overlapping word ‘road’, whereas neither Sense 1 nor Sense 2 share any words with the sentence. The authors use Wikipedia articles as dictionary definitions. So, to disambiguate the word ‘cone’ in the sentence ‘I found a cone on the road outside my house.’, articles such as ‘Cone (category theory)’[6], ‘Cone (formal languages)’[7] and ‘Cone (linear algebra)’[8] would be used in place of dictionary definitions.

[5] These definitions come from the Oxford Online Dictionary and represented a truncated list of the full list of definitions for ‘cone’. The full list can be seen here: `http://www.oxforddictionaries.com/definition/english/cone`

[6] `https://en.wikipedia.org/wiki/Cone_(category_theory)`

[7] `https://en.wikipedia.org/wiki/Cone_(formal_languages)`

[8] `https://en.wikipedia.org/wiki/Cone_(linear_algebra)`

Feature-based

This method represents the context in which an ambiguous phrase appears in an unseen text and contexts in which it appears as a link in Wikipedia articles in terms of specific characteristics called *features*. A single grouping containing a particular value for each feature is called a *feature set*. The features Mihalcea and Csomai extract are as follows:

```
‘‘the current word and its part-of-speech, a local context of three words
to the left and right of the ambiguous word, the parts-of-speech of the
surrounding words, and a global context implemented through sense specific
keywords determined as a list of at most five words occurring at least three
times in the contexts defining a certain word sense.’’
```

A feature set is collected for every context (paragraph) where the keyphrase appears as a link in a Wikipedia article, along with the title of the article to which it links. Suppose for example the phrase ‘cone’ appeared as a link in 100 contexts[9], linking to ‘Cone (category theory)’ 30 times, ‘Cone (formal languages)’ 50 times, and ‘Cone (linear algebra)’ 20 times. This yields 100 feature sets, each of which is associated with a single article title, called its *class*. A feature set *f_unseen* is also created for the context in which the term appears in the unseen text. The task is to determine the class for *f_unseen* by comparing its feature sets with the 100 feature sets that have been collected. Mihalcea and Csomai employ a Naïve Bayes classifier to make this determination.

Naïve Bayes classifiers calculate the probability of a feature set referring to each of the possible classes. They then choose the class that has the highest probability. In the example given in the previous paragraph, this means calculating the probability that a phrase’s feature set refers to each of ‘Cone (category theory)’, ‘Cone (formal languages)’ and ‘Cone (linear algebra)’ articles and choosing the one whose probability is highest. The probability that a phrase’s feature set *f_unseen* refers to a particular Wikipedia article A is given by the following formula:

$$P(A|f\_unseen) = P(A) \cdot P(f\_unseen|A) \tag{B.5}$$

where:

- P$(A|f_unseen)$ is the probability that the phrase, represented by the feature set *f_unseen*, has the class A

[9] Note that 100 contexts does not necessarily mean 100 articles, as a phrase can appear as a link in the same article more than once.

- P(A) is the probability of selecting a random feature set whose class is A. For example P('Cone (category theory)') = 0.3 (i.e. 30 divided by 100).

- P($f\_unseen|A$) is the probability that the class A would produce *f_unseen*. This is calculated by first determining the probability that each individual feature value in *f_unseen* would produce A. Suppose a particular feature value appears in a total of 40 feature sets; 20 feature sets whose class is 'Cone (linear algebra)' and 20 feature sets whose class is 'Cone (category theory)'. This feature value's probability of producing 'Cone (linear algebra)' is 0.5, as is its probability of producing 'Cone (category theory)'. The feature value's probability of producing 'Cone (formal languages)' is 0. The individual probabilities for each feature value in *f_unseen* are multiplied together to give P($f\_unseen|A$).

Mihalcea and Csomai evaluate the feature-based and knowledge-based disambiguation algorithms on the same dataset they used to evaluate the keyphrase extraction algorithms. The authors perform a comparison against two baselines, one which assigns a sense at random and one which chooses the most common sense for the ambiguous phrase i.e. the article to which it most frequently links in Wikipedia. The authors also compare against an approach that combines the knowledge-based and feature-based methods. This approach only selects a particular sense if it is returned by both methods. The evaluation metrics used were again precision, recall and F-measure. In this case, precision refers to the ratio of the number of phrases correctly disambiguated by the method to the total number of disambiguations made by the method. Recall refers to the proportion of the manually annotated keyphrases that were correctly disambiguated by the method. The results of Mihalcea and Csomai's evaluation can be seen in Table B.6.

Method	Precision	Recall	F-measure
Random sense	63.84	56.90	60.17
Most frequent sense	87.03	77.57	82.02
Knowledge-based	80.63	71.86	75.99
Feature-based learning	92.91	**83.10**	**87.73**
Combined	**94.33**	70.51	80.69

Table B.6: Mihalcea and Csomai's results from comparing knowledge- and feature-based WSD algorithms.

Appendix C

The Schulze Method

The Schulze Method is a rank-aggregation method that was created for the purposes of determining the winning candidate in single-winner elections[1] [111]. Per Schulze: "Today (May 2014), the proposed method is used by more than 60 organizations with more than 100,000 eligible members in total. Therefore, the proposed method is more wide-spread than all other Condorcet-consistent single-winner election methods together."

In the Schulze method there is a collection A of candidates. Different rankings of this list are called *Strict Weak Orders* on A. A strict weak order is a binary relation $\succ$ defined on a collection of items where $a \succ b$ means 'a is ranked higher than b'. A strict weak order has the following properties:

- Asymmetricity - This property is satisfied if only one of the following statements is true:
  - $a \succ b$ (i.e. a is ranked higher than b)
  - $b \succ a$ (i.e. b is ranked higher than a)
  - $a \approx b$ (i.e. neither $a \succ b$ nor $b \succ a$)
- Irreflexivity - This property states that $a \approx a$. In other words, a cannot be ranked higher or lower than itself.
- Transitivity - This property states that if $a \succ b$ and $b \succ c$, then $a \succ c$. That is, if a is ranked higher than b, and b is ranked higher than c, then a must be ranked higher than c.
- Negative Transitivity - This property states that if $a \not\succ b$ and $b \not\succ c$ then $a \not\succ$c. That is, if a is not ranked higher (i.e. is ranked lower or equal) than b, and b is not ranked higher than c, then a cannot be ranked higher than c.

[1] A single-winner election is an election in which only one candidate can win.

The Schulze Method takes a collection of strict weak orders, called *ballots* as input and returns as output:

1. A non-empty set of winners.

2. A *Strict Partial Order* on A. A strict partial order is a relation that is asymmetric, irreflexive and transitive. This strict partial order is the aggregated ranking of the items in A.

The first step in the Schulze Method is to compute a pairwise ranking matrix *N*. *N[a,b]* represents the number of voters who preferred a to b. Table C.1 shows an example collection of ballots. Table C.2 shows the associated pairwise ranking matrix. From Table C.2, *N[a,b]* $= 8$, because a was preferred to b in 8 ballots. N defines *paths* between candidates. A path between two candidates *C(1)* and *C(n)* is a sequence of 2 or more candidates. A path consists of *links* between candidates *C(i)* and *C(i+1)* such that :

1. $C(i) \neq C(i{+}1)$. In words, the candidate at position i must be different from the candidate at position *i+1*.

2. $N[C(i), C(i+1)] \succ N[C(i+1), C(i)]$ or $N[C(i), C(i+1)] \approx N[C(i+1), C(i)]$. In words, the candidate in position i must have either a greater than or equal number of preferred ballots to the candidate in position *i+1*.

The strength of a path is given by the strength of a path's weakest link i.e. the value of the lowest cell *N[C(i),C(i+1)]* in the path.

The next step in the Schulze Method is to calculate the strongest path between each candidate. For example, one path from a to b is: *N[a,c]* (14), *N[c,b]* (15). *N[a,c]* is greater than *N[c,a]* and *N[c,b]* is greater than *N[b,c]*. The strength of this path is 14. The other path between a and b is: *N[a,c]* (14), *N[c,d]* (12), *N[d,b]* (19). The strength of this path is 12. Thus, the strength of the strongest path between a and b is :maximum(14, 12) = 14. Table C.3 shows the pairwise matrix M of the strongest paths between each of the candidates. A candidate W is declared the winner if and only if, for every other candidate C, *M[W,C]* is greater than *M[C, W]*. So, in Table 5.4, d is the winner. The remaining candidates are ranked based on the number of pairwise comparisons they win. The ranking of the remaining candidates from Table 5.4 is: $a \succ (b, c)$. In this ranking, b and c tie because they each win one pairwise comparison.

Number of ballots	**Ranking**
8	$a \succ c \succ d \succ b$
2	$b \succ a \succ d \succ c$
4	$c \succ d \succ b \succ a$
4	$d \succ b \succ a \succ c$
3	$d \succ c \succ b \succ a$

Table C.1: Example ballots.

	N[*,a]	N[*,b]	N[*,c]	N[*,d]
N[a,*]	—	8	14	10
N[b,*]	13	—	6	2
N[c,*]	7	15	—	12
N[d,*]	11	19	9	—

Table C.2: A pairwise comparison matrix for Table C1.

	N[*,a]	N[*,b]	N[*,c]	N[*,d]
N[a,*]	—	14	14	12
N[b,*]	13	—	13	12
N[c,*]	13	15	—	12
N[d,*]	13	19	13	—

Table C.3: A pairwise comparison matrix of strongest paths for Table C.2.

Appendix D

Rank-Order Centroids

Barron and Barrett conduct a comparison between three direct weight elicitation methods [120]. The authors' comparison represents items using the *Multiattribute Value Model (MAV).* This model is concerned with the way in which a single alternative is chosen from a list of alternatives. Each alternative in the list has the same attributes, but different values for each attribute. The Multiattribute Value for an alternative a with n attributes is given by:

$$MAV_a = \sum_{i=1}^{n} w_i \cdot v_{ai} \tag{D.1}$$

where v_{ai} is the ath alternative's value for the ith attribute and w_i is the weight assigned to the ith attribute. The sum of all weights equals 1. After the MAV values for each alternative have been computed, the item with the highest MAV is chosen. The weight elicitation task consists of automatically determining w_i for every attribute i, given only the rank-order of the attributes. In other words, the task is to convert each attribute's rank into a numerical value, its weight.

Barron and Barrett compare three different weight elicitation methods. Each of these is described below.

Rank-sum (RS)

In RS, weights correspond to ranks, normalised by dividing by the sum of the ranks. The formula is as follows:

$$w_i(RS) = \frac{n+1-i}{\sum_{j=1}^{n} j} \tag{D.2}$$

	Rank-Sum	Rank-Reciprocal	Rank-Order Centroids
w1	0.4000	0.4800	0.5208
w2	0.3000	0.2400	0.2708
w3	0.2000	0.1600	0.1458
w4	0.1000	0.1200	0.0625

Table D.1: The Rank-Sum, Rank-Reciprocal and Rank-Order Centroid weights for four ranked items.

Rank-reciprocal (RR)

RR is similar to RS but is based on the reciprocal of the ranks. The formula is below:

$$w_i(RR) = \frac{1/i}{\sum_{j=1}^{n} \frac{1}{j}} \tag{D.3}$$

Rank-order centroids (ROC)

In ROC, weights are computed from the vertices of a *simplex*. A simplex is the "simplest geometrical figure in a given dimension, so the line, triangle and tetrahedron are the simplices in one, two and three dimensions" [116]. The number of dimensions for the simplex is the same as the number of attributes i.e. n. The vertices of the simplex are $e_1 = (1, 0, ..., 0), e_2 = (\frac{1}{2}, \frac{1}{2}, 0, ...0), e_3 = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0, ..., 0), e_n = (\frac{1}{n}, \frac{1}{n}, ..., \frac{1}{n})$. The weights for each attribute are represented by the *centroid* of the simplex. The centroid is computed by calculating the mean of the values for each attribute, as follows:

$$w_i(ROC) = \frac{1}{n} \sum_{j=i}^{n} \frac{1}{j} \tag{D.4}$$

The RR, RS and ROC values for four ranked items can be seen in Table D.1.

Barron and Barrett compare the RS, RR and ROC methods by performing a computer simulation. The first step in this simulation is to generate a set of 'true' weights. Recall that the ranking of attributes is known beforehand, therefore it is known which attribute should have the highest weight, which should have the next highest etc. A single collection of weights can thus be generated by sampling from the range 0-1 for each attribute, while ensuring that the sum of all the weights is one. Such a weight combination will henceforth be referred to as TRUE. Each of the RS, RR and ROC methods are then employed to generate weights. The chosen alternative using the TRUE weights (i.e. the alternative with the highest MAV) is designated the 'correct' alternative. The RS, RR and ROC methods are examined by comparing the alternatives they select with the correct alternative. This process can be summarised as follows:

1. Generate a random vector representing TRUE weights.
2. Generate a list of m alternatives, each with random values for each attribute. The values are scaled so that the smallest value is 0 and the largest value is 1.
3. Calculate MAV_a for each of the m alternatives, separately for each of the TRUE, ROC, RS, RR weights. As a baseline, the authors also use a weighting method EW that assigns equal weights to every attribute.
4. Compare the selections under TRUE with the selections under the other weighting methods.

m is varied between 3,6,9,12 and 15. For each m, the number of attributes n is varied between 5,10,15,20 and 25. Values come from four different kinds of distribution:

1. Uniformly generated, uncorrelated - Uniformly generated means that values are selected randomly from the range [0,1] with equal probability. Uncorrelated means that a change in the value of one attribute doesn't affect the value of another.
2. Normally generated, uncorrelated - Normally generated means that values are selected randomly from a normal distribution over the range [0,1].
3. Normally generated, negatively correlated. Negatively correlated means that if an attribute changes in one direction, another attribute changes in the opposite direction. For example if one attribute increases, another decreases.
4. Normally generated, positively correlated. Positively correlated means that if an attribute changes in a particular direction, another attribute changes in the same direction.

An example of the uniform (green) and normal (red line) distributions can be seen in Figure D.1. The y-axis indicates the probability with which a value will be chosen. With the uniform distribution, each value has the same likelihood of being selected. With the normal distribution, some values are far more likely than others. Each combination of m and n for each of the four possible distributions yields 100 different design elements in total (i.e. 5 x 5 x 4). For each of these, 100 random value matrices are created, as are 100 sets of random weights. This gives 10,000 trials per design element. The following metrics are used to evaluate each method:

- Hit Rate - The proportion of all cases in which an examined method (ROC, RS, RR or EW) selects the same alternative as TRUE.

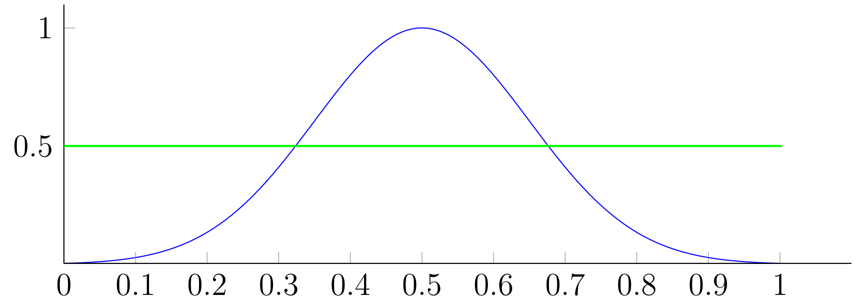


Figure D.1: An example of a normal (blue) and uniform (green) distribution over the range [0,1].

- Average Value Loss - The average absolute difference in MAV for the alternative selected by TRUE weights and the alternative selected by an examined method (ROC, RS, RR or EW).

- Average Proportion of Maximum Value Range Achieved - Let $V_{TRUE(a)}$ denote the MAV for alternative a calculated using the TRUE weights. For each trial determine the indices $a_{max:TRUE}$ and $a_{min:TRUE}$ corresponding to the largest and smallest MAVs generated using the TRUE weights. Then determine the index of the selected alternative using an examined method, for example ROC. The Proportion of Maximum Value Range Achieved (PMVRA) is given by:

  $$\frac{V_{TRUE}(a_{max:ROC}) - V_{TRUE}(a_{min:TRUE})}{V_{TRUE}(a_{max:TRUE}) - V_{TRUE}(a_{min:TRUE})} \tag{D.5}$$

  PMVRA is a comparison of what the gap between the alternative selected by ROC and the lowest TRUE alternative actually is to what the gap should be. PMVRA is also calculated for the other methods.

Each of these metrics is averaged across the 10,000 trials for each of the 100 design elements.

Value Type	Hit Rate	PMVRA	Value Loss
Uniform, independent	22	25	25
Normal, independent	25	25	25
Normal, negative correlation	24	24	25
Normal, positive correlation	25	25	25

Table D.2: Bannon and Barrett's results from comparing the Rank-Reciprocal and Rank-Order Centroid weights.

Barron and Barrett find that the Rank-Order Centroid method outperforms all the others by a large margin. In the vast majority of cases, the ordering was $ROC > RR > RS > EW$. Table D.2 shows the number of times this order held. For the 'Uniform, independent' Hit Rate cell , RS performed better than RR when there were 3 attributes and when the number of alternatives was 10,15 and 20. For the 'Normal, negative correlation' Hit Rate cell, RS performed better than ROC when there were 25 alternatives and the number of attributes was 3. For the 'Normal, negative correlation' PMVRA cell, RS performed better than RR when there were 20 alternatives and the number of attributes was 3.

These results show that the ROC method is the most effective in generating weights for a ranked list.

Appendix E

Vacancy Categories

Bodleian Libraries	Centre for Health Service Economics and Organisation	Centre for Tropical Medicine and Global Health	Department of Biochemistry	Department of Chemistry	Department of Engineering Science	Department of Oncology
Department of Paediatrics	Department of Physics	Department of Politics and International Relations (DPIR)	Faculty of Classics	Faculty of History	Faculty of Law	Humanities Division
Kellogg College	Mathematical Institute	Medical Sciences Division	NOT FOUND (refers to three vacancies for which no category was provided.)	Nuffield Department of Clinical Medicine	Nuffield Department of Population Health	Nuffield Department of Primary Care Health Sciences
Personnel and Related Services	Radcliffe Department of Medicine	Radcliffe School of Medicine	School of Geography and the Environment	Social Sciences Division	University Administration and Services (UAS)	University of Oxford

Table E.1: Job vacancy categories

Appendix F

MovieLens Genres

Action	Adventure	Animation	Children
Comedy	Crime	Documentary	Drama
Fantasy	Film-Noir	Horror	Musical
Mystery	Romance	Thriller	Unknown
Sci-fi	War		

Table F.1: MovieLens genres

Appendix G

Music dataset genres

Alternative	Blues	Classical	Comedy
Country	Dance	Easy Listening	Electronic
Experimental	Folk	Funk	Hip Hop
Holiday	Indie	Industrial	Inspirational
Instrumental	Jazz	Karaoke	Metal
New Age	Pop	Punk	Rap
Reggae	R&B	Rock	Singer-songwriter
Soul	Spoken Word	Vocal	World

Table G.1: Music dataset genres

Bibliography

[3] P. Brusilovsky, V. P. Wade, and O. Conlan, “Architecture solutions for e-learning systems,” in. Idea Group Inc, 2007, pp. 243–261.

[4] P. Brusilovsky, “Methods and techniques of adaptive hypermedia,” *User modeling and user-adapted interaction*, vol. 6, no. 2, pp. 87–129, 1996.

[5] B. Steichen, A. O’Connor, and V. Wade, “Personalisation in the wild: Providing personalisation across semantic, social and open-web resources,” in *HT’11, Proceedings of the 22nd ACM Conference on Hypertext and Hypermedia, Eindhoven, The Netherlands, June 6-9, 2011*, pp. 73–82.

[6] B. Hecht, J. Teevan, M. R. Morris, and D. J. Liebling, “Searchbuddies: Bringing search engines into the conversation,” in *Proceedings of the Sixth International Conference on Weblogs and Social Media, Dublin, Ireland, June 4-7, 2012.*

[7] F. Abel, Q. Gao, G. Houben, and K. Tao, “Analyzing user modeling on twitter for personalized news recommendations,” in *User Modeling, Adaption and Personalization - 19th International Conference,UMAP 2011, Girona, Spain, July 11-15, 2011. Proceedings*, pp. 1–12.

[8] R. Malone, *Structuring unstructured data*, `http://www.forbes.com/2007/04/04/teradata-solution-software-biz-logistics-cx_rm_0405data.html`, 2007.

[9] W. Balke, “Introduction to information extraction: Basic notions and current trends,” *Datenbank-Spektrum*, vol. 12, no. 2, pp. 81–88, 2012.

[10] O. Etzioni, M. Banko, S. Soderland, and D. S. Weld, “Open information extraction from the web,” *Communications of the ACM*, vol. 51, no. 12, pp. 68–74, 2008.

[11] S. Auer, C. Bizer, G. K. andJens Lehmann andRichard Cyganiak, and Z. G. Ives, “Dbpedia: A nucleus for a web of open data,” in *The Semantic Web*, vol. 4825, 2007, pp. 722–735.

[12] J. Vassileva, J. Blustein, L. Aroyo, and S. K. D’Mello, Eds., *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization, UMAP 2016, Halifax, NS, Canada, July 13 - 17, 2016*, ACM.

[13] Y. Huang, M. Yudelson, S. Han, D. He, and P. Brusilovsky, "A framework for dynamic knowledge modeling in textbook-based learning," in *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization, UMAP 2016, Halifax, NS, Canada, July 13 - 17, 2016*, 2016, pp. 141–150.

[14] J. Herzig, G. Feigenblat, M. Shmueli-Scheuer, D. Konopnicki, and A. Rafaeli, "Predicting customer satisfaction in customer support conversations in social media using affective features," in *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization, UMAP 2016, Halifax, NS, Canada, July 13 - 17, 2016*, 2016, pp. 115–119.

[15] K. Oyibo, Y. S. Ali, and J. Vassileva, "Gender difference in the credibility perception of mobile websites: A mixed method approach," in *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization, UMAP 2016, Halifax, NS, Canada, July 13 - 17, 2016*, 2016, pp. 75–84.

[16] G. Piao and J. G. Breslin, "Analyzing aggregated semantics-enabled user modeling on google+ and twitter for personalized link recommendations," in *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization, UMAP 2016, Halifax, NS, Canada, July 13 - 17, 2016*, 2016, pp. 105–109.

[17] M. Rokicki, E. Herder, T. Kusmierczyk, and C. Trattner, "Plate and prejudice: Gender differences in online cooking," in *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization, UMAP 2016, Halifax, NS, Canada, July 13 - 17*, pp. 207–215.

[18] A. Kangasrääsiö, Y. Chen, D. Glowacka, and S. Kaski, "Interactive modeling of concept drift and errors in relevance feedback," in *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization, UMAP 2016, Halifax, NS, Canada, July 13 - 17, 2016*, 2016, pp. 185–193.

[19] I. Andjelkovic, D. Parra, and J. O'Donovan, "Moodplay: Interactive mood-based music discovery and recommendation," in *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization, UMAP 2016, Halifax, NS, Canada, July 13 - 17, 2016*, 2016, pp. 275–279.

[20] D. Jannach, I. Kamehkhosh, and G. Bonnin, "Biases in automated music playlist generation: A comparison of next-track recommending techniques," in *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization, UMAP 2016, Halifax, NS, Canada, July 13 - 17, 2016*, 2016, pp. 281–285.

[21] M. Michelson and S. A. Macskassy, "Discovering users' topics of interest on twitter: A first look," in *Proceedings of the Fourth Workshop on Analytics for Noisy Unstructured Text Data, Toronto, Ontario, Canada, October 26th, 2010 (in conjunction with CIKM)*, pp. 73–80.

[22] P. Kapanipathi, P. Jain, C. Venkatramani, and A. P. Sheth, "User interests identification on twitter using a hierarchical knowledge base," in *The Semantic Web: Trends and Challenges - 11th International Conference, ESWC 2014, Anissaras, Crete, Greece, May 25-29, 2014. Proceedings*, pp. 99–113.

[23] C. Chiarcos, S. Hellmann, and S. Nordhoff, "Introduction and overview," in *Linked Data in Linguistics - Representing and Connecting Language Data and Language Metadata.* 2012, pp. 1–12.

[24] L. Bentivogli, P. Forner, B. Magnini, and E. Pianta, "Revising the wordnet domains hierarchy: Semantics, coverage and balancing," in *Proceedings of the Workshop on Multilingual Linguistic Ressources*, ser. MLR '04, 2004, pp. 101–108.

[25] S. Alanazi, J. Goulding, and D. McAuley, "Cross-system recommendation: User-modelling via social media versus self-declared preferences," in *Proceedings of the 27th ACM Conference on Hypertext and Social Media, 2016, Halifax, NS, Canada, July 10-13*, 2016, pp. 183–188.

[26] C. Paris, P. Thomas, and S. Wan, "Differences in language and style between two social media communities," in *Proceedings of the Sixth International Conference on Weblogs and Social Media, Dublin, Ireland, June 4-7, 2012*, 2012.

[27] J. Arguello, F. Diaz, J. Callan, and J. Crespo, "Sources of evidence for vertical selection," in *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Boston, MA, USA, July 19-23, 2009*, pp. 315–322.

[28] F. Abel, N. Henze, E. Herder, and D. Krause, "Interweaving public user profiles on the web," in *User Modeling, Adaptation, and Personalization, 18th International Conference, UMAP 2010, Big Island, HI, USA, June 20-24, 2010. Proceedings*, pp. 16–27.

[29] F. Orlandi, J. G. Breslin, and A. Passant, "Aggregated, interoperable and multi-domain user profiles for the social web," in *I-SEMANTICS 2012 - 8th International Conference on Semantic Systems, I-SEMANTICS '12, Graz, Austria, September 5-7, 2012*, pp. 41–48.

[30] X. Guo, H. Jerbi, and M. P. O'Mahony, "An analysis framework for content-based job recommendation," in *22nd International Conference on Case-Based Reasoning (ICCBR*, 2014.

[31] C. Hauff and G. Houben, "Deriving knowledge profiles from twitter," in *Towards Ubiquitous Learning - 6th European Conference of Technology Enhanced Learning, EC-TEL 2011, Palermo, Italy, September 20-23, 2011. Proceedings*, pp. 139–152.

[32] B. Flyvbjerg, "Five misunderstandings about case-study research," *Qualitative Inquiry*, pp. 219–245, 2006.

[33] B. Adams and K. Janowicz, "On the geo-indicativeness of non-georeferenced text," in *Proceedings of the Sixth International Conference on Weblogs and Social Media, Dublin, Ireland, June 4-7, 2012.*

[34] F. Abel, C. Hauff, G. Houben, R. Stronkman, and K. Tao, "Semantics + filtering + search = twitcident. exploring information in social web streams," in *23rd ACM Conference on Hypertext and Social Media, HT '12, Milwaukee, WI, USA, June 25-28, 2012.*

[35] J. Ding, L. Gravano, and N. Shivakumar, "Computing geographical scopes of web resources," in *VLDB 2000, Proceedings of 26th International Conference on Very Large Data Bases, September 10-14, 2000, Cairo, Egypt*, pp. 545–556.

[36] V. R. Martínez, L. E. P. Estrada, F. Iacobelli, S. S. Bojórquez, and V. M. González, "Semi-supervised approach to named entity recognition in spanish applied to a real-world conversational system," in *Pattern Recognition - 7th Mexican Conference, MCPR 2015, Mexico City, Mexico, June 24-27, 2015, Proceedings*, pp. 224–235.

[37] B. Heitmann, M. Dabrowski, A. Passant, C. Hayes, and K. Griffin, "Personalisation of social web services in the enterprise using spreading activation for multi-source, cross-domain recommendations," in *Intelligent Web Services Meet Social Computing, Papers from the 2012 AAAI Spring Symposium, Palo Alto, California, USA, March 26-28, 2012.*

[38] J. B. Alonso, C. Havasi, and H. Lieberman, "Perspectivespace: Opinion modeling with dimensionality reduction," in *User Modeling, Adaptation, and Personalization, 17th International Conference, UMAP 2009, formerly UM and AH, Trento, Italy, June 22-26, 2009. Proceedings*, pp. 162–172.

[39] G. A. Miller, "Wordnet: A lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[40] G. Semeraro, P. Basile, M. Degemmis, and P. Lops, "Content-based recommendation services for personalized digital libraries," in *Digital Libraries: Research and Development, First International DELOS Conference, Pisa, Italy, February 13-14, 2007, Revised Selected Papers*, pp. 77–86.

[41] P. Bellekens, L. Aroyo, G. Houben, A. Kaptein, and K. van der Sluijs, "Semantics-based framework for personalized access to TV content: The ifanzy use case," in *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007.*, pp. 887–894.

[42] G. Klyne, J. J. Carroll, and B. McBride, "Resource description framework (rdf): Concepts and abstract syntax," World Wide Web Consortium, Tech. Rep., Feb. 2004.

[43] A. Miles and S. Bechhofer, *Skos simple knowledge organization system reference*, `https://www.w3.org/TR/skos-reference/`, 2009.

[44] D. Yang, T. Nie, D. S. andGe Yu, and Y. Kou, "Personalized web search with user geographic and temporal preferences," in *Web Technologies and Applications - 13th Asia-Pacific Web Conference, APWeb 2011, Beijing, China, April 18-20, 2011. Proceedings*, pp. 95–106.

[45] K. Tao, F. Abel, Q. Gao, and G. Houben, "TUMS: twitter-based user modeling service," in *The Semantic Web: ESWC 2011 Workshops - ESWC 2011 Workshops, Heraklion, Greece, May 29-30, 2011, Revised Selected Papers*, pp. 269–283.

[46] Q. Gao, F. Abel, G. Houben, and Y. Yu, "A comparative study of users' microblogging behavior on sina weibo and twitter," in *User Modeling, Adaptation, and Personalization - 20th International Conference, UMAP 2012, Montreal, Canada, July 16-20, 2012. Proceedings*, pp. 88–101.

[47] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing and Management.*, vol. 24, no. 5, pp. 513–523, 1988.

[48] E. Gilbert and K. Karahalios, "Predicting tie strength with social media," in *Proceedings of the 27th International Conference on Human Factors in Computing Systems, CHI 2009, Boston, MA, USA, April 4-9, 2009*, pp. 211–220.

[49] F. Abel, Q. Gao, G. Houben, and K. Tao, "Semantic enrichment of twitter posts for user profile construction on the social web," in *The Semanic Web: Research and Applications - 8th Extended Semantic Web Conference, ESWC 2011, Heraklion, Crete, Greece, May 29 - June 2, 2011, Proceedings, Part II*, pp. 375–389.

[50] Y. Ding and J. Jiang, "Extracting interest tags from twitter user biographies," in *Information Retrieval Technology - 10th Asia Information Retrieval Societies Conference, AIRS 2014, Kuching, Malaysia, December 3-5, 2014. Proceedings*, pp. 268–279.

[51] C. Lioma and I. Ounis, "Examining the content load of part of speech blocks for information retrieval," in *ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July.*

[52] T. D. Smedt and W. Daelemans, "Pattern for python," *Journal of Machine Learning Research*, vol. 13, pp. 2063–2067, 2012.

[53] W. Daelemans, J. Zavrel, P. Berck, and S. Gillis, "MBT: A memory-based part of speech tagger-generator," *CoRR*, vol. cmp-lg/9607012, 1996.

[54] W. Daelemans, S. Buchholz, and J. Veenstra, "Memory-based shallow parsing," in *Proceedings of the 1999 Workshop on Computational Natural Language Learning, CoNLL-99, Held in cooperation with EACL'99, Bergen, Norway, June 12, 1999*, pp. 53–60.

[55] P. Brusilovsky and E. Millán, "User models for adaptive hypermedia and adaptive educational systems," in *The Adaptive Web, Methods and Strategies of Web Personalization*, 2007, pp. 3–53.

[56] J. Stan, V. Do, and P. Maret, "Semantic user interaction profiles for better people recommendation," in *International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2011, Kaohsiung, Taiwan, 25-27 July 2011*, pp. 434–437.

[57] Q. Gao, F. Abel, and G. Houben, "Genius: Generic user modeling library for the social semantic web," in *The Semantic Web - Joint International Semantic Technology Conference, JIST 2011, Hangzhou, China, December 4-7, 2011. Proceedings*, pp. 160–175.

[58] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer, "Dbpedia spotlight: Shedding light on the web of documents," in *Proceedings the 7th International Conference on Semantic Systems, I-SEMANTICS 2011, Graz, Austria, September 7-9, 2011*, 2011, pp. 1–8.

[59] P. N. Mendes, A. Passant, P. Kapanipathi, and A. P. Sheth, "Linked open social signals," in *2010 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2010, Toronto, Canada, August 31 - September 3, 2010, Main Conference Proceedings*, pp. 224–231.

[60] B. Heap, A. Krzywicki, W. Wobcke, M. Bain, and P. Compton, "Combining career progression and profile matching in a job recommender system," in *PRICAI 2014: Trends in Artificial Intelligence - 13th Pacific Rim International Conference on Artificial Intelligence, Gold Coast, QLD, Australia, December 1-5, 2014. Proceedings*, pp. 396–408.

[61] G. Piao and J. G. Breslin, "Analyzing mooc entries of professionals on linkedin for user modeling and personalized mooc recommendations," in *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, Halifax, Nova Scotia, Canada, pp. 291–292.

[62] C. Li, F. Wang, Y. Yang, Z. Li, and X. Zhang, "Exploring social network information for solving cold start in product recommendation," in *Web Information Systems Engineering - WISE 2015 - 16th International Conference, Miami, FL, USA, November 1-3, 2015, Proceedings, Part II*, 2015, pp. 276–283.

[63] H. Ko, I. Ko, T. Kim, D. Lee, and S. J. Hyun, "Identifying user interests from online social networks by using semantic clusters generated from linked data," in *Current Trends in Web Engineering - ICWE 2013 International Workshops ComposableWeb, QWE, MDWE, DMSSW, EMotions, CSE, SSN, and PhD Symposium, Aalborg, Denmark, July 8-12, 2013. Revised Selected Papers*, 2013, pp. 302–309.

[64] J. C. dos Reis, R. Bonacin, and M. C. C. Baranauskas, "Beyond the social search: Personalizing the semantic search in social networks," in *Online Communities and Social Computing - 4th International Conference, OCSC 2011, Held as Part of HCI International 2011, Orlando, FL, USA, July 9-14, 2011. Proceedings*, 2011, pp. 345–354.

[65] M. D. Choudhury, S. Counts, and M. Gamon, "Not all moods are created equal! exploring human emotional states in social media," in *Proceedings of the Sixth International Conference on Weblogs and Social Media, Dublin, Ireland, June 4-7, 2012*, 2012.

[66] A. Mislove, B. Viswanath, P. K. Gummadi, and P. Druschel, "You are who you know: Inferring user profiles in online social networks," in *Proceedings of the Third International Conference on Web Search and Web Data Mining, WSDM 2010, New York, NY, USA, February 4-6, 2010*, 2010, pp. 251–260.

[67] Statista, *Number of monthly active twitter users worldwide from 1st quarter 2010 to 1st quarter 2018 (in millions)*, `https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/`, 2018.

[68] L. Hong, G. Convertino, and E. H. Chi, "Language matters in twitter: A large scale study," in *Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011*, 2011.

[69] BusinessDictionary, *What is customer support?* http://www.businessdictionary.com/definition/customer-support.html.

[70] B. Diirr, R. M. de Araujo, and C. Cappelli, "Talking about public service processes," in *Electronic Participation - Third IFIP WG 8.5 International Conference, ePart 2011, Delft, The Netherlands, August 29 - September 1, 2011. Proceedings*, pp. 252–261.

[71] H. Weigand, P. Johannesson, B. Andersson, J. J. Arachchige, and M. Bergholtz, "Management services - A framework for design," in *Advanced Information Systems Engineering - 23rd International Conference, CAiSE 2011, London, UK, June 20-24, 2011. Proceedings*, pp. 582–596.

[72] J. Hise and D. W. Massey, "Applying the ignatian pedagogical paradigm to the creation of an accounting ethics course," *Journal of Business Ethics*, vol. 96, no. 3, pp. 453–465, 2010.

[73] M. Naaman, J. Boase, and C. Lai, "Is it really about me?: Message content in social awareness streams," in *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work, CSCW 2010, Savannah, Georgia, USA, February 6-10, 2010*, 2010, pp. 189–192.

[74] C. Honeycutt and S. C. Herring, "Beyond microblogging: Conversation and collaboration via twitter," in *42st Hawaii International International Conference on Systems Science (HICSS-42 2009), Proceedings (CD-ROM and online), 5-8 January 2009, Waikoloa, Big Island, HI, USA*, 2009, pp. 1–10.

[75] S. Wunsch-Vincent and G. Vickery, "Participative web: User-created content," Organisation for Economic Co-operation and Development, Tech. Rep., Mar. 2007.

[76] Y. Ma, Y. Zeng, X. Ren, and N. Zhong, "User interests modeling based on multi-source personal information fusion and semantic reasoning," in *Active Media Technology - 7th International Conference, AMT 2011, Lanzhou, China, September 7-9, 2011. Proceedings*, pp. 195–205.

[77] H. Mao, X. Shuai, and A. Kapadia, "Loose tweets: An analysis of privacy leaks on twitter," in *Proceedings of the 10th annual ACM workshop on Privacy in the electronic society, WPES 2011, Chicago, IL, USA, October 17, 2011*, 2011, pp. 1–12.

[78] G. Rizzo and R. Troncy, "Nerd: Evaluating named entity recognition tools in the web of data," in *ISWC 2011, Workshop on Web Scale Knowledge Extraction (WEKEX'11), October 23-27, 2011, Bonn, Germany*, 2011.

[79] J. Giles, "Internet encyclopaedias go head to head," *Nature*, vol. 438, pp. 900–901, Dec. 2005.

[80] M. Zhang, B. J. Jansen, and A. Chowdhury, “Business engagement on twitter: A path analysis,” *Electronic Markets*, vol. 21, no. 3, pp. 161–175, 2011.

[81] H. Alharthi and D. Inkpen, “Content-based recommender system enriched with wordnet synsets,” in *Computational Linguistics and Intelligent Text Processing - 16th International Conference, CICLing 2015, Cairo, Egypt, April 14-20, 2015, Proceedings, Part II*, pp. 295–308.

[82] M. Grassi and F. Piazza, “Towards an RDF encoding of conceptnet,” in *Advances in Neural Networks - ISNN 2011 - 8th International Symposium on Neural Networks, ISNN 2011, Guilin, China, May 29-June 1, 2011, Proceedings, Part III*, pp. 558–565.

[83] J. Bullas, *5 insights into the latest social media facts, figures and statistics*, `http://www.jeffbullas.com/2013/07/04/5-insights-into-the-latest-social-media-facts-figures-and-statistics/`, 2013.

[84] A. Gonzalez-Agirre, G. Rigau, and M. Castillo, “A graph-based method to improve wordnet domains,” in *13th International Conference on Computational Linguistics and Intelligent Text Processing, New Delhi, India, March 11-17, 2012, Proceedings,Part I*, pp. 17–28.

[85] P. University, *What is wordnet?* `http://wordnet.princeton.edu/`, 2016.

[86] L. Borin and M. Forsberg, “Swesaurus; or, the frankenstein approach to wordnet construction,” in *Proceedings of the Seventh Global Wordnet Conference*, 2014, pp. 215–223.

[87] P. D. Turney, “Coherent keyphrase extraction via web mining,” in *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, 2003, pp. 434–439.

[88] E. Riloff, “Understanding language understanding,” in, A. Ram and K. Moorman, Eds., MIT Press, 1999, ch. Information Extraction As a Stepping Stone Toward Story Understanding, pp. 435–460.

[89] Y. Chang, R. Z. andSrihari Reddy, and Y. Liu, “Detecting multilingual and multiregional query intent in web search,” in *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2011, San Francisco, California, USA, August 7-11, 2011.*

[90] J. Vosecky, D. Jiang, K. W. Leung, and W. Ng, “Dynamic multi-faceted topic discovery in twitter,” in *22nd ACM International Conference on Information and Knowledge Management, CIKM’13, San Francisco, CA, USA, October 27 - November 1, 2013*, pp. 879–884.

[91] L. Ermakova and J. Mothe, “IRIT at INEX 2013: Tweet contextualization track,” in *Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013.*.

[92] B. Magnini, C. Strapparava, G. Pezzulo, and A. M. Gliozzo, “The role of domain information in word sense disambiguation,” *Natural Language Engineering*, vol. 8, no. 4, pp. 359–373, 2002.

[93] J. Schler, M. Koppel, S. Argamon, and J. W. Pennebaker, “Effects of age and gender on blogging,” in *Computational Approaches to Analyzing Weblogs, Papers from the 2006 AAAI Spring Symposium, Technical Report SS-06-03, Stanford, California, USA, March 27-29, 2006*, pp. 199–205.

[94] O. U. Press, *Rt this: Oup dictionary team monitors twitterer’s tweets*, `http://blog.oup.com/2009/06/oxford-twitter/`, 2009.

[95] H. Kwak, C. Lee, H. Park, and S. B. Moon, “What is twitter, a social network or a news media?” In *Proceedings of the 19th International Conference on World Wide Web, 2010, Raleigh, North Carolina, USA, April 26-30*, pp. 591–600.

[96] R. MacMillan, *Michael kinsley and the length of newspaper articles*, `http://blogs.reuters.com/mediafile/2010/01/05/michael-kinsley-and-the-length-of-newspaper-articles/`, 2010.

[97] S. P. Ponzetto and R. Navigli, “Knowledge-rich word sense disambiguation rivaling supervised systems,” in *ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden*, pp. 1522–1531.

[98] D. McCarthy and J. A. Carroll, “Disambiguating nouns, verbs, and adjectives using automatically acquired selectional preferences,” *Computational Linguistics*, vol. 29, no. 4, pp. 639–654,

[99] E. Agirre and A. Soroa, “Personalizing pagerank for word sense disambiguation,” in *EACL 2009, 12th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, Athens, Greece, March 30 - April 3, 2009*, 2009, pp. 33–41.

[100] P. Gomes, F. C. Pereira, P. Paiva, N. Seco, P. Carreiro, J. L. Ferreira, and C. Bento, “Noun sense disambiguation with wordnet for software design retrieval,” in *Proceedings of the 16th Canadian Society for Computational Studies of Intelligence Conference on Advances in Artificial Intelligence*, ser. AI’03, Halifax, Canada, 2003, pp. 537–543.

[101] J. Bullas, *25 linkedin facts and statistics you need to share*, `http://www.jeffbullas.com/2014/12/02/25-linkedin-facts-and-statistics-you-need-to-share/`, 2014.

[102] F. A. Zamal, W. Liu, and D. Ruths, “Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors,” in *Proceedings of the Sixth International Conference on Weblogs and Social Media, Dublin, Ireland, June 4-7, 2012.*

[103] J. Roberts, *Typical twitter user is a young woman with an iphone & 208 followers*, `https://gigaom.com/2012/10/10/the-typical-twitter-user-is-a-young-woman-with-an-iphone-and-208-followers/`, 2012.

[104] Z. Yu, P. Wang, X. Du, J. Cui, and T. Xu, “A timeline-based algorithm for personalized tag recommendation,” in *Web Information Systems Engineering - WISE 2010 Workshops - WISE 2010 International Symposium WISS, and International Workshops CISE, MBC, Hong Kong, China, December 12-14, 2010, Revised Selected Papers*, pp. 378–389.

[105] T. Plumbaum, S. Wu, E. W. D. Luca, and S. Albayrak, “User modeling for the social semantic web,” in *Proceedings of the second Workshop on Semantic Personalized Information Management: Retrieval and Recommendation 2011, Bonn, Germany, October 24*, pp. 78–89.

[106] R. Baeza-Yates and B. Ribeiro-Neto, “Modern information retrieval: The concepts and technology behind search,” in. Pearson Education Limited, 2011, ch. 3, pp. 68–75.

[107] Y. Zeng, E. Zhou, Y. Wang, X. Ren, Y. Qin, Z. Huang, and N. Zhong, “Research interests: Their dynamics, structures and applications in unifying search and reasoning,” *J. Intell. Inf. Syst.*, vol. 37, no. 1, pp. 65–88, 2011.

[108] H. Wu, J. He, Y. Pei, and X. Long, “Finding research community in collaboration network with expertise profiling,” in *Advanced Intelligent Computing Theories and Applications, 6th International Conference on Intelligent Computing, ICIC 2010, Changsha, China, August 18-21, 2010. Proceedings*, 2010, pp. 337–344.

[109] M. Taylor and D. Richards, “Discovering areas of expertise from publication data,” in *Knowledge Acquisition: Approaches, Algorithms and Applications: Pacific Rim Knowledge Acquisition Workshop, PKAW 2008, Hanoi, Vietnam, December 15-16, 2008, Revised Selected Papers*. 2009, pp. 218–230.

[110] A. Moro, A. Raganato, and R. Navigli, “Entity linking meets word sense disambiguation: A unified approach,” *TACL*, vol. 2, pp. 231–244, 2014.

[111] M. Schulze, “A new monotonic, clone-independent, reversal symmetric, and condorcet-consistent single-winner election method,” *Social Choice and Welfare*, vol. 36, no. 2, pp. 267–303, 2014.

[112] D. Greene and P. Cunningham, “Practical solutions to the problem of diagonal dominance in kernel document clustering,” in *Proc. 23rd International Conference on Machine learning (ICML’06)*, 2006, pp. 377–384.

[113] F. M. Harper and J. A. Konstan, “The movielens datasets: History and context,” *ACM Transactions on Interactive Intelligent Systems*, vol. 5, no. 4, p. 19, 2016.

[114] S. Oramas, V. C. Ostuni, T. D. Noia, X. Serra, and E. D. Sciascio, “Sound and music recommendation with knowledge graphs,” *ACM Transactions on Intelligent Systems and Technology*, vol. 8, no. 2, 21:1–21:21, 2016.

[115] G. Salton, A. Wong, and C. S. Yang, “A vector space model for automatic indexing,” *Commun. ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[116] C. Clapham and J. Nicholson, “The concise oxford dictionary of mathematics: Fourth edition,” in. Wiley-Interscience, 2000, pp. 364–372.

[117] A. Dattolo, F. Ferrara, and C. Tasso, “Supporting personalized user concept spaces and recommendations for a publication sharing system,” in *User Modeling, Adaptation, and Personalization, 17th International Conference, UMAP 2009, Trento, Italy, June 22-26. Proceedings*, pp. 325–330.

[118] T. D. Noia, R. Mirizzi, V. C. Ostuni, D. Romito, and M. Zanker, “Linked open data to support content-based recommender systems,” in *I-SEMANTICS 2012 - 8th International Conference on Semantic Systems, I-SEMANTICS ’12, Graz, Austria, September 5-7, 2012*, 2012, pp. 1–8.

[119] D. M. Buede, “The engineering design of systems,” in. Wiley-Interscience, 2000, ch. 13, pp. 364–372.

[120] F. H. Barron and B. E. Barrett, “Decision quality using ranked attribute weights,” *Management Science*, vol. 42, no. 11, pp. 1515–1523, 1996.

[121] B. S. Ahn and K. S. Park, “Comparing methods for multiattribute decision making with ordinal weights,” *Computers & Operations Research*, vol. 35, no. 5, pp. 1660–1670, 2008.

[122] R. Real and J. M. Vargas, ““‘” *Systematic Biology*, vol. 45, no. 3, pp. 380–385, 1996.

[123] L. R. Dice, “Measures of the amount of ecologic association between species,” *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[124] D. Diez, C. Barr, and M. Çetinkaya-Rundel, *OpenIntro Statistics*. OpenIntro, Incorporated, 2015.

[125] D. H. Park, H. K. Kim, I. Y. Choi, and J. K. Kim, “A literature review and classification of recommender systems research,” *Expert Systems with Applications*, vol. 39, no. 11, pp. 10 059–10 072, 2012.

[126] Y. K. Lee and H. T. Ng, "An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation," in *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, ser. EMNLP '02, 2002, pp. 41–48.

[127] B. Magnini and G. Cavaglia, "Integrating subject field codes into wordnet," in *Proceedings of the Second International Conference on Language Resources and Evaluation, LREC 2000, 31 May - June 2, 2000, Athens, Greece.*

[128] R. Mihalcea and A. Csomai, "Wikify!: Linking documents to encyclopedic knowledge," in *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, Lisbon, Portugal, November 6-10, 2007*, pp. 233–242.

[129] M. Lesk, "Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone," in *Proceedings of the 5th Annual International Conference on Systems Documentation*, 1986, pp. 24–26.