An Evaluation and Enhancement of Densitometric Fragmentation for Content Slicing Reuse

Killian Levacher
Trinity College Dublin
killian.levacher@cscc.tcd.ie

Seamus Lawless
Trinity College Dublin
seamus.lawless@scss.tcd.ie

Vincent Wade
Trinity College Dublin
vincent.wade@scss.tcd.ie

ABSTRACT
Content slicing addresses the need of adaptive systems to reuse open corpus material by converting it into re-composable information objects. However this conversion is highly dependent upon the ability to correctly fragment pages into structurally sound atomic pieces. A recently suggested approach to fragmentation, which relies on densitometric page representation, claims to achieve high accuracy and time performance. Although it has been well received within the research community, a full evaluation of this approach and identification of strengths and weaknesses across a range of characteristics hasn't been performed. This paper proposes an independent evaluation of the approach with respect to granularity control, accuracy, time performance, content diversity and linguistic dependency. Moreover, this paper also provides a significant contribution to address important weaknesses discovered during the analysis, in order to improve the suitability and impact of the original algorithm within the context of content slicing.

Categories and Subject Descriptors
H.3.3 [**Information Systems**]: Information Search and Retrieval;

Keywords
Analysis, Densitometric, Fragmentation, Open-Corpus

1. INTRODUCTION
Adaptive Hypermedia Systems (AHS) have traditionally attempted to deliver dynamically adapted and personalised presentations to users through the sequencing of re-composable pieces of information [7]. However their inherent reliance upon bespoke, proprietary content represents a major obstacle to their widespread adoption. The adaptivity that an AHS can deliver is restricted due to a lack of sufficient content in terms of volume, granularity, style and meta-data. In response to this challenge, open corpus content (multilingual information accessible on the WWW and digital repositories) is increasingly incorporated within AHS [2]. However, web content returned by traditional Information Retrieval (IR) systems is not directly suitable for re-composition. It is only available as "one size fits all" content, with limited control over granularity, format or meta-data, which are critical requirements for re-composability [4]. As a result, it is extremely difficult for AHS to incorporate externally authored content when generating adaptive offerings. A good example of what could be achieved, if open corpus resources were fully available as re-composable content objects, is the Personal Multilingual Customer Care system developed by Steichen et al [6]. This system leverages both corporate and user generated resources available in the wild, by integrating and re-composing these resources into single coherent presentations, meeting user specific unique information needs (based on expertise and/or prior interaction). However, such systems do so by manually converting targeted resources into re-composable objects using manually crafted, content specific rule-based algorithms. These techniques do not scale if the entire open web is to be targeted as a potential resource [5]. Slicing techniques address these limitations by automatically converting open corpus resources into re-composable information objects called slices, tailored for consumption by individual AHS. However as open corpus content is very heterogeneous and generally contains unnecessary noise such as navigation bars, advertisements etc, the process of slicing (described in section 2) is highly dependent upon the ability to correctly fragment heterogeneous pages into structurally sound atomic pieces. A recently suggested approach to fragmentation, which relies on densitometric page representation [3] claims to achieve high accuracy and time performance. Although it has been well received within the research community, it appears, to the author's knowledge, that a full evaluation of this approach and identification of strengths and weaknesses across a range of characteristics critical for content slicing, such as granularity control, time performance, content type or multilingual dependency, hasn't been performed. Since fragmentation represents a decisive step within the context of a content slicing system, an evaluation of this approach determining its suitability for slicing purposes is required.

Contribution: This paper hence proposes i) an independent evaluation of the approach with original results in relation to granularity control, content diversity, linguistic dependency, time performance and accuracy. Moreover, ii) a significant contribution to address some weaknesses discovered during the analysis, and increase the suitability and impact of this fragmentation technique within the context of content slicing systems is provided by this paper.

2. OPEN CORPUS SLICING
In order to convert open-corpus content into slices, a slicer must automatically process a diverse and large range of documents at very high speed. For the purpose of this research, the slicer framework in question [5] consists of a pipeline divided in four distinct components: 1) Harvesting**:** the first module consists in identifying and harvesting open-corpus content from various large repositories using focused crawling techniques [4]; 2) Fragmentation: Standard sections of harvested content are then fragmented into structurally sound atomic pieces (such as menus, advertisements and main articles). This phase is critical since

CIKM'12, October 29-November 2, 2012, Maui, HI, USA.

"there is an inverse relationship between the potential reusability of content and its granularity" [4]. Maximizing the reuse potential of a news article from a previously published document, along with original menus and advertisements, is far less adequate than reusing the article alone, de-contextualized from its original setting and at various levels of granularity; 3) Semantic Annotations: once fragmented, each resulting fragment is annotated with semantic labels using pre-selected algorithms; 4) Slice Creation: the fourth step finally combines the resulting fragments and annotations into standardized slice units, ready for delivery to third party slice consumers.

Fragmentation Requirements: Hence, within the context of a slicer, a fragmenter must provide the ability to process very large amount of heterogeneous pages (in the region of tens of millions) in i) multiple languages and ii) of various content types (forum pages, product pages, news pages…) with iii) a strong control over granularity of fragments produced, at iv) a very high speed, and v) on demand. Since the set of pages prior to slicing is unknown and hence the number of possible DOM layout patterns is infinite, fragmentation should also occur without the need for interpreting the vi) meaning or vii) structure of tags within a page.

3. DENSITOMETRIC FRAGMENTATION

As web pages have evolved over time, content fragmentation has become an increasingly difficult task to perform. Besides the wide variety of pages available, each individual page itself now contains an increasing set of heterogeneous elements such as user comments, advertisements, snippets previews etc. Among the wide variety of algorithms designed for the purpose of content fragmentation, a densitometric approach [3] appeared to be the most promising with respect to slicing requirements. Its ability to fragment documents regardless of the meaning or structure of xml tags used, allows it to process virtually any xml-based document (requirements vi and vii). Moreover, as it considers words as mere tokens, this would theoretically make it language and content type agnostic (requirements i and ii). Furthermore, as its fragmentation process requires no rendering, this would significantly reduce any computational costs, hence increases time performance (requirement iv). The aim of this section is to present the fundamental concepts related to densitometric fragmentation required for the purpose of understanding the subsequent analysis. For a more complete description, the reader is referred to the original paper [3]. Within the context of densitometric fragmentations, DOM tag tree leaf nodes are converted into a 1 dimensional array of text density values. The text density $\rho(\tau_x)$ (Equation 1a) of a tag τ_x, is defined as the ratio between the number of tokens and the number of lines within τ_x.

$$\text{a) } \rho(\tau_x) = \frac{\sum Tokens_{\tau_x}}{\sum Lines_{\tau_x}} \quad \text{b) } \Delta_\rho(i,i+1) = \frac{\left|\rho(i) - \rho(i+1)\right|}{\max\left(\rho(i),\rho(i+1)\right)} \quad (1)$$

A line is defined as a group of successive characters, with a total character number equal to an arbitrary word wrapping value ω_x. Tags containing only one line of text are assigned a text density value equal to the number of tokens it possesses. Although a line might appear as an arbitrary notion, sharp changes in text density correspond relatively well to desired fragmentation. Densitometric fragmentation therefore consists in identifying such variations in text density and correlating them with fragment boundaries. All densitometric fragmentation algorithms proceed in detecting these variations by considering each leaf tag text density value as one atomic block (or fragment). Each page is hence converted into a single block array. Fragmentation algorithms subsequently iterate through this array by selecting individual blocks and then fusing them together into larger compounded blocks to form a new block array. These iterations are performed multiple times, by fusing compounded blocks together, until final fragments are created. Densitometric algorithms differ upon the criteria selected for fusion as well as how adjacent blocks are selected prior to fusion.

Plain fusion for instance, only considers pairs of adjacent blocks (block window size of 2) at a time. If the text density difference Δ_ρ (Equation 1b) of the pair is smaller than an arbitrary threshold value $\nu_{\max}$, the blocks are fused and the resulting compounded block produced is compared to the next and so on.

Smooth fusion extends the previous algorithm by considering a blocks predecessor and successor block text densities (block window size of 3). If these text densities are identical and higher than the blocks own density, all three blocks are fused.

Rule-based fusions on the other hand, attempt to augment the previous algorithms by taking into account the meaning of specific tags (titles, tables etc…) in order to infer structural composition of pages. Whenever such a tag is encountered, a block fusion or block gap is performed regardless of densitometric values. However, as taking into account the meaning of tags violates requirement v), rule based fusions represent the least desirable densitometric variation within the context of a content slicing pipeline.

4. DENSITOMETRIC ANALYSIS

4.1 Evaluation Objectives

The purpose of this evaluation was to provide a critical analysis of densitometric algorithms with respect to characteristics (section 2) currently unavailable within the literature and critical for slicer pipelines. These concerns are encapsulated within the following interrogations. Can densitometric algorithms provide control over the resulting granularity of fragments produced (section 4.2)? Is the precision of these algorithms independent of content type or language considered for fragmentation (section 4.3, 4.4)? Finally, can such a fragmentation approach occur with similar time performances for various chosen granularities (section 4.5)? In this evaluation, i) plain fusion, ii) smooth fusion, iii) pure rule-based and iv)smooth rule-based fusion variations were implemented and tested for the purpose of this analysis. Since smooth fusions achieved very similar results with respect to plain and pure rule-based variations, these algorithms are omitted in this paper for clarity purposes. All algorithms were evaluated over a parallel multilingual corpus of approximately 20,000 pages acquired from our commercial partner Microsoft. This corpus consisted of MS Office manuals in four different languages (English, French, German and Spanish).

Adjusted Random Index (ARI): Accuracies were measured using the standard ARI metric [7]. This index measures the similarities between two clustering methods by determining the number of agreements within two vectors, using values ranging from 0 (no agreement) to 1 (perfect agreement).

4.2 Granularity Analysis

Granularity variation of fragments was measured with respect to threshold $\nu_{\max}$. A granularity percentage index was computed for

each page fragmented by calculating the difference in blocks between pre and post fusion and normalizing each value according to the total number of blocks present prior to fragmentation. Values close to 100 hence correspond to very large granularities, while those closer to 0 represent small granularities. Results obtained for plain fusion algorithm (executed over the English subset of the corpus consisting of approximately 5000 pages), depicted a direct linear increase in granularity across ν_{max} values ranging between 10 and 90. While Non rule-based variations provided granularities with a difference of 76% between smallest and largest fragments produced, rule-based variation only provided a difference of 15%. This suggests non-rule-based fusions provide a much higher range of granularity with respect to their rule-based counter part. However, while rule-based fusions provide a stable standard deviation, non-rule-based fusions possess a standard deviation of up to 30% for ν_{max} values ranging between 0.4 and 0.7. This suggests a trade off between the range of resulting fragment sizes available and the predictability of this granularity. In other words, a rule-based approach to fragmentation will offer a very small range of granularity however most pages will possess similar granularity, while non-rule based fragmentations provide a wider range of granularity with less predictability for each fragment.

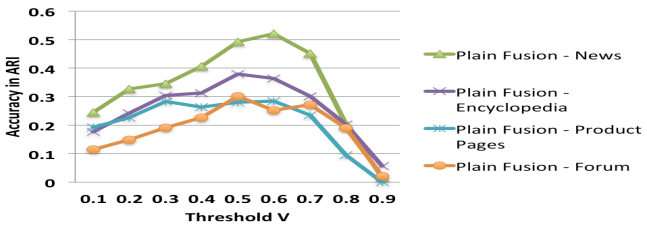

Figure 1 Plain Fusion Accuracy Across Content Types

4.3 Content Type Accuracy Performance

The following analysis investigated any possible content type dependencies of densitometric fragmentations. Within this experiment, content types refer to 4 general page types namely, Encyclopedias, Forum, Product and News pages. Five human annotators manually annotated a sample corpus of 250 pages (randomly selected and available online) from each content type. In order to provide a comparison baseline with other studies, the news set consisted of a subset of the corpus used in the original paper [3]. Figure 1 presents the results obtained using plain fusion fragmentation across content types. As can be seen, results depict an intuitive ranking of content type accuracy by degree of editorial control. News and encyclopedia pages achieve accuracy values close to 60%, however product and in particular forum pages depict very poor accuracies in the low 20%. Subsequent examination of forum pages revealed that forum posts contained both the actual content of the post with post menus, each possessing respectively very high and low text densities, which densitometric fusion algorithms are designed to separate. This results in the title and menus of each post being separated from the post content which explains the poor performance of forum content over all fusion versions. Product page densitometric values depict the same pattern. Hence, this experiment strongly suggests a densitometric algorithm dependency upon the type of content being segmented, with higher accuracies obtained for pages containing fragments with continuous regions of similar high or low densities rather than fragments with alternating high/low densities. A lot of care must hence be taken prior to selecting this algorithm with respect to the type of content envisaged to process.

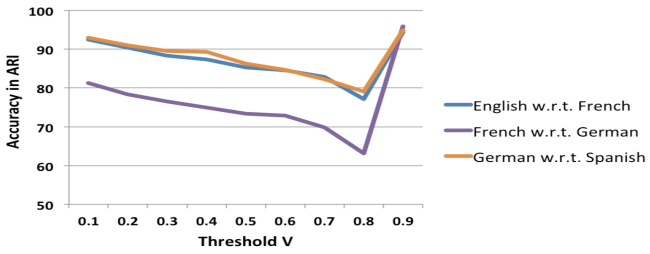

Figure 2 Plain Fusion Across Languages

4.4 Multilingual Analysis

As pointed out in section 1, open corpus content is by nature multilingual. The ability therefore to predict the similarity and accuracy of fragments produced from a set of multilingual parallel documents represents an important slicing requirement. Hence, in this experiment the full 20,000 pages of the multilingual (English, French, German, Spanish) parallel corpora were fragmented using the plain fusion approach. Within this corpus, each English file possesses an xml structurally identical twin. The only difference for each parallel file set is the linguistic content within each tag. Hence, any fragmentation variations between equivalent files can only be due to linguistic differences. Figure 2, presents the results obtained by comparing each fragment produced for every parallel file set combination. Language combinations not presented within the graph were omitted for clarity purposes as they achieved very similar results to the German w.r.t Spanish combination. Results suggest fragment similarities, although very high, will decrease in average by 2% for every 0.1 increase in ν_{max} for all $\nu_{max} < 0.8$. Past this value, the accuracy increases again very fast, resulting in very high accuracy sensitivity for $\nu_{max} > 0.8$. A standard deviation of only 2% was also measured across all language combinations, which suggests a high predictability in the fragment similarity expected for given parameters. Hence, although a densitometric fusion approach to content fragmentation is linguistically agnostic, a predictable decrease in similarity across resulting language fragments must be taken into account while fragmenting open corpus content. Finally, results also reveal how the French/German language combination consistently portrayed accuracies 10% lower than others. Further experiments are currently investigating whether this difference is caused by word length distribution dependency between languages.

4.5 Greedy Algorithm & Time Performance

The lower chart in Figure 3 depicts a time performance analysis carried out upon the plain fusion algorithm. As can be noticed, an increase in time necessary to process pages, for large granularity requirements ($\nu_{max} > 0.6$), was discovered. This discovery unfortunately makes the algorithm very unattractive within the context of slicers since a slicer requires the ability to fragment open corpus pages at different levels of granularity and at high speed. A new greedy densitometric algorithm (algorithm 1) was hence designed in order to stabilize time performance across granularities by reducing the number of iterations necessary through the block array. This algorithm replaces a fixed block window size with a variable one with the aim of making most fusions occur in early iterations and stabilize time performances for high ν_{max} values. A greedy behavior drives this window

expansion, including additional adjacent blocks based on local densitometric value variations. This algorithm also differs from existing approaches by using a variable threshold value ν_{max} automatically adjusted with respect to local regions of a page based on densitometric values of blocks currently selected within the window. The assumption is that, in addition to driving the window expansion, adjusting fusion threshold values within specific regions of pages should increase the fragmentation accuracy. As one can observe in Figure 3, the greedy window-expanding algorithm reduces significantly the average time needed for block fusion per page for high threshold values ν_{max}. And although it is non-rule based, this algorithm depicts very similar time performance behaviors as its rule based alternative with a performance increase of 56% in average with respect to plain fusion and up to 89% improvement for threshold values $\nu_{max} \approx 0.9$. Finally, slight accuracy improvements with both Encyclopedia and News content types can be observed (for $\nu_{max} > 0.3$ and $\nu_{max} < 0.8$).

ALGORITHM 1: GREEDY FUSION

Main Function: Input: block_array b[b1,..,bn], default_
Ouput: fragmented page as compound block array b[]
begin
| fusedBlocks = true
| **while** fusedBlocks
| | fusedBlocks = false, index=0
| | **while** index < b[].size
| | | windowSize=**Fusable_Window_Size**(index,b[])
| | | **if** windowSize >0 then
| | | | **fuse**(index, windowSize, b[])
| | | | fusedBlocks = true
| | | **end**
| | | index++;
| | **end**
| **end**
end

Function: Fusable_Window_Size
Input: index, block_array b[b1,b2,..bn], default_ ν_{max}
Output: window size to fuse
begin
| ν_{max}_values.*add*(default_ ν_{max})
| windowExpanded =true; blockMerged =0
| greedyIndex = index+blockMerged
| **while** windowExpanded and (greedyIndex+1)<= b[].size
| | = **avg**(ν_{max} _values)
| | windowExpanded = false
| | densitoDifference = Δ(b[greedyIndex],b[greedyIndex+1])
| | **if** densitoDifference < then
| | | ν_{max} _values.add(densitoDifference)
| | | blockMerged ++
| | | windowExpanded = true
| | | greedyIndex = index+blockMerged
| | **end**
| **end**
end return blockMergeD

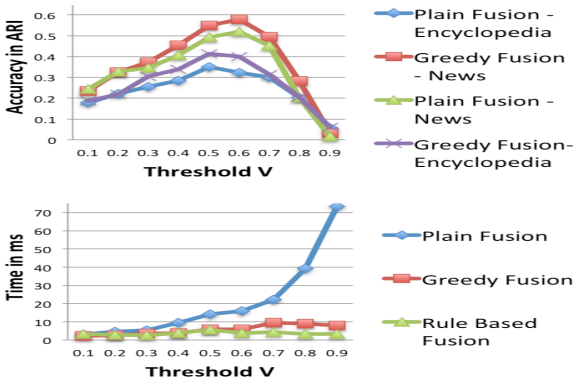

Figure 3 Greedy vs Plain Fusion

5. CONCLUSION AND FUTURE WORK

The analysis performed upon densitometric fragmentation algorithms confirms the suitability of this approach with respect to its ability to fragment without the need for interpreting the meaning or structure of tags within pages. It additionally provides the ability to fragment pages across languages with a predictable control over a wide range of granularities and cross-language accuracies. However this approach is highly content type dependent, with higher accuracies achieved for news and encyclopedia content. A significant time performance decrease for high granularity values was also discovered which makes this algorithm unattractive within the context of a slicing system. Despite this drawback, a new greedy fusion algorithm, provides a significant increase in time performances, which stabilizes time performance across granularities and provides higher accuracies.

6. ACKNOWLEDGMENTS

This research is supported by the SFI (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie).

7. REFERENCES

[1] Hearst, M.A. TextTiling : Segmenting Text into Multi-paragraph Subtopic Passages. *Journal of Computational Linguistics* (1997).

[2] Henze, N. and Nejdl, W. Adaptivity in the KBS Hyperbook. *WWW'99: Workshop on Adaptive Systems and User Modelling*, (1999)

[3] Kohlschütter, C. and Nejdl, W. A Densitometric Approach to Web Page Segmentation. *CIKM'08: Proc. of the 17th int. conf. on Information and knowledge management*, (2008),.

[4] Lawless, S. Leveraging Content from Open Corpus Sources for Technology Enhanced Learning 2009.

[5] Levacher, K., Wade, V et.al. Providing Customized Reuse of Open-Web Resources for Adaptive Hypermedia. *23rd Conf. on Hypertext and Social Media*, (2012).

[6] Steichen, B. and Wade, V. Providing Personalisation across Semantic , Social and Open-Web Resources. *22nd int. conf. on Hypertext and Hypermedia*, (2011)

[7] Strehl, A. Cluster Ensembles – A Knowledge Reuse Framework for Combining Multiple Partitions. *Journal of Machine Learning Research 3*, (2002)