Term Frequency Dynamics in Collaborative Articles

Sérgio Nunes[†]
ssn@fe.up.pt

Cristina Ribeiro[†,‡]
mcr@fe.up.pt

Gabriel David[†,‡]
gtd@fe.up.pt

[‡]INESC-Porto
[†]DEI, Faculdade de Engenharia, Universidade do Porto
Rua Dr. Roberto Frias, s/n
4200-465 Porto, Portugal

ABSTRACT

Documents on the World Wide Web are dynamic entities. Mainstream information retrieval systems and techniques are primarily focused on the latest version a document, generally ignoring its evolution over time. In this work, we study the term frequency dynamics in web documents over their lifespan. We use the Wikipedia as a document collection because it is a broad and public resource and, more important, because it provides access to the complete revision history of each document. We investigate the progression of similarity values over two projection variables, namely revision order and revision date. Based on this investigation we find that term frequency in encyclopedic documents – i.e. comprehensive and focused on a single topic – exhibits a rapid and steady progression towards the document's current version. The content in early versions quickly becomes very similar to the present version of the document.

Categories and Subject Descriptors

I.7.5 [**Document and Text Processing**]: Document Capture—*Document analysis*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Measurement, Experimentation

Keywords

Document Dynamics, Term Frequency, Wikipedia

1. INTRODUCTION

Documents on the World Wide Web are dynamic entities. Contrary to other communication media, where different versions of the same document are treated as separate documents (e.g. print), on the web this distinction is less obvious. Web document are rarely static, exhibiting a high degree and rate of change. To understand the internal dynamics of web documents over time, we observe and measure the changes between different versions of the same document. We focus our study on a particular type of web documents – collaboratively written documents, more specifically Wikipedia articles. The size, scope and popularity of Wikipedia, together with the fact that the complete revision history of its articles is available, make this collection a unique and appropriate resource for this investigation.

We start by modeling each version of a document as a term frequency vector and use the cosine similarity measure to quantify the differences between past versions and the current version of each article. To investigate the progression of similarity values we consider two different projections axes, namely revision order and revision date. Depending on the projection variable used, the similarity curves exhibit different shapes. We also contrast the internal dynamics of high quality documents – as determined by the Wikipedia community – with a sample of ordinary documents.

DocEng2010, September 21–24, 2010, Manchester, United Kingdom.

2. RELATED WORK

First works in web dynamics were focused in studying the behavior of the web as a whole, with the primary goal of optimizing the crawling of web pages (e.g. define revisitation patterns). Early reference work in this area includes the papers by Ntoulas et al [4] and Fetterly et al [3].

In recent studies, more attention has been given to the internal dynamics of web documents for retrieval tasks. Adar et al [1] conducted a detailed observation of a large collection of popular web pages and were able to clearly distinguish between stable and dynamic content. The stable part of a document is defined as the content that remains the same over time. Using *change curves* plots, the authors showed that the stable content of a page tends to stabilize after a short period of time. The work by Elsas and Dumais [2] is one of the first to analyze the temporal dynamics of document content to improve relevance ranking. Using a collection of top ranked web documents, the authors establish a relationship between content change patterns and document relevance. They observe that highly relevant documents are more likely to change than documents in general, both in terms of frequency and degree. Both these works present evidence that supports the importance of the temporal dimension in information retrieval tasks. Our work is different since it focuses on the complete lifespan of web documents, from inception to its current version, as opposed to the observation of subsequent changes in popular documents. In other words, while previous works are focused on the changes made to existing documents, we try to measure the internal dynamics of a web document since its creation. Moreover,

we look at typical documents from a wiki, instead of popular pages with frequently updated content (e.g. news sites, portals, forums).

Our investigation has similarities with the work published by Thomas and Sheth [5] on the content dynamics of Wikipedia articles. These authors model each revision to an article as a "TF-IDF vector" and use a cosine similarity measure to evaluate the convergence of content. We explore a similar idea to model changes between revisions. However, these authors are focused in a classification problem – distinguish high quality articles from lower quality articles by looking at content evolution. They found no statistically significant difference between both types of articles in terms of edit history. Our work has a different context, it is focused on the measurement and characterization of the internal dynamics of a document's content for information retrieval tasks. Moreover, while these authors use an absolute revision-based timeline to observe the evolution of content, we use normalized projections to overcome the problem of article comparability.

3. DOCUMENT COLLECTION

We use the English version of Wikipedia[1] to assemble a collection of documents for analysis. We select all articles currently classified as *featured article* (i.e. belonging to *Category:Featured articles*) and a parallel set of random articles obtained using the *random article* feature available in Wikipedia[2]. The featured article category contains articles identified by the community as high quality documents, frequently singled out in Wikipedia's frontpage. The most significant difference between these two groups of documents is in the total number of revisions, with featured articles having a significantly higher number of total revisions. Table 1 summarizes the distribution of the number of revisions in both sets. For each set of documents we present the value for 1st and 3rd quartile, the median and the mean number of revisions. It is worth noting that, to avoid sampling documents with a very small number of total revisions, we discard all random articles with less than 50 revisions.

Table 1: Document collection overview.

	Articles	Number of Revisions			
		1st Q.	Median	Mean	3rd Q.
Featured	2,710	348.5	645.5	1,363	1,534
Random	2,430	65	100	226	188

4. TERM FREQUENCY DYNAMICS

Our motivation for this work is to understand how term frequency evolves over time within documents. To carry out this investigation, we model each version of a document as a *term frequency (tf) vector* and use the cosine of the angle between the two vectors to quantify the similarity of two document versions. With this approach, similarity values vary between -1 for opposite vectors and 1 for identical vectors, with 0 for orthogonal vectors. We remove all wiki-markup and stop words from each version of the document before assembling the *tf* vectors.

To observe how content evolves within a document, we compare each version of a document with its current version.

[1] `http://en.wikipedia.org`

[2] `http://en.wikipedia.org/wiki/Special:Random`

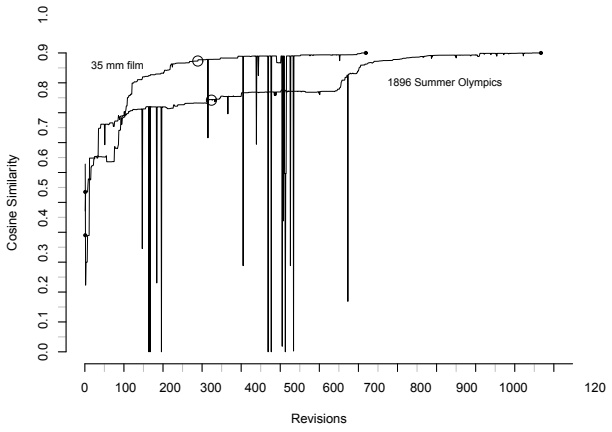

Figure 1: Similarity by revision order in two Wikipedia articles.

As an example, consider Figure 1 which shows the evolution of this similarity measure by revision for two featured articles. The Wikipedia article about "35 mm film" currently has slightly over 700 revisions, while the article about the "1896 Summer Olympics" has approximately 1200 revisions. As can be seen in the figure, the *similarity profile* of both articles steadily converges to 1, although at different paces. This is the expected result – each revision made to an article moves it closer to its current version. Worth of notice is the fact that content similarity tends to evolve quite rapidly, reaching high levels after a relatively small number of revisions. For instance, in the "35 mm film" the cosine similarity between the version at revision 200 and the latest version of the article (over revision 700) is over 0.9. The abrupt drops observed in both profiles are due to vandalism, a well-known problem in Wikipedia. Finally, we have highlighted with a circle the revision where each article was added to the *featured articles* category.

Although this figure reveals quite similar trends in two different articles, it also shows that it is difficult to compare the similarity evolution of articles with a distinct number of revisions. For this reason, we propose the normalization of the horizontal axis based on a quantile discretization approach. This way we are able to observe, side by side, articles with a different number of revisions and obtain a comparative picture of the internal dynamics of content across a broad group of documents. We explore two different projection variables for a normalized horizontal axis – revision order and revision date. The following sections describe each approach.

4.1 Discretizing by Revision Order

We establish 25-quantiles for each article's revision history and extract the content for each of the 25 bins. For instance, in an article with a total of 50 revisions, we first extract revision 2 (1st bin), then revision 4 (2nd bin), and so forth. The content found in each of the 25 revisions (bins) is compared with the content available in revision 50 (the current version of the article). Figure 2 depicts a series of boxplots, each one summarizing the similarity values found between each bin and the current version of the document, for all the 2,710 featured articles. Note that the *x-axis* is represented in a $[0, 1]$ scale for consistency. This picture shows a clear

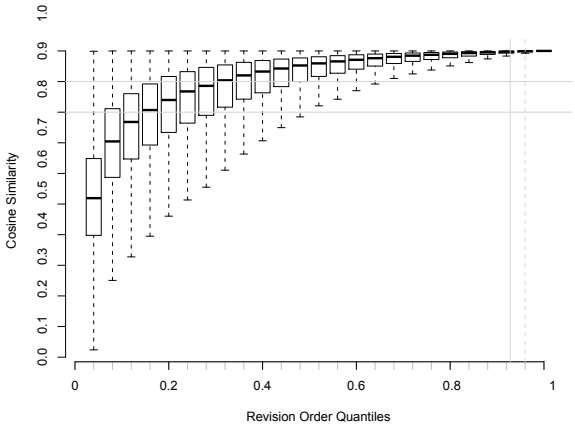

Figure 2: Boxplots of cosine similarity for featured articles, discretized by revision order.

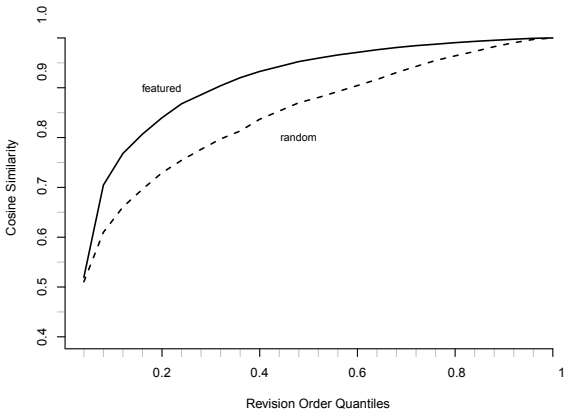

Figure 4: Median cosine similarity of featured and random articles, discretized by revision order.

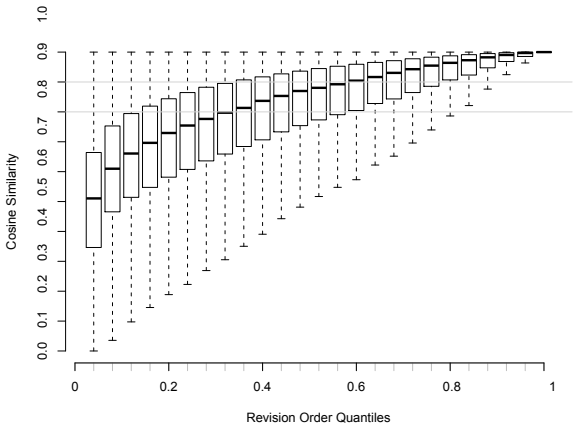

Figure 3: Boxplots of cosine similarity for random articles, discretized by revision order.

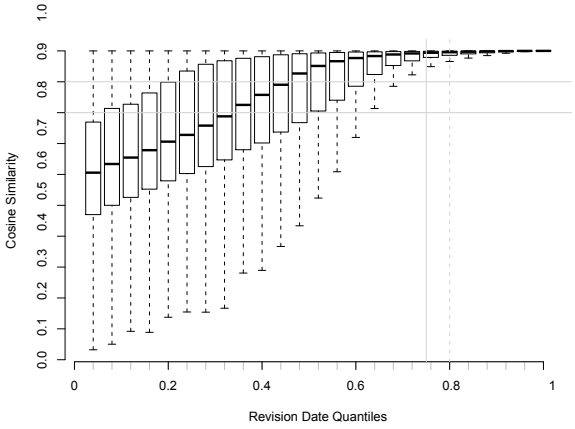

Figure 5: Boxplots of cosine similarity for featured articles, discretized by revision date.

pattern about the evolution of content similarity over revision order. As can be seen with the help of the horizontal lines added, the median similarity is over 0.8 since the 4th bin, i.e. at 16% ($\frac{4}{25}$) of the revision history of an article. Moreover, halfway the revision history of featured articles, the 1st quartile of similarity values is higher than 0.9. In other words, in more than 75% of all featured articles the intermediate revision is already very similarity to the current version. Subsequent changes to an article have very little impact in a document's term frequency vector. This figure clearly shows the rapid progress of content similarity in featured articles. When looking at a comparable plot based on the random collection of articles mentioned before, a somewhat different picture appears as seen in Figure 3. Although the similarity values also move consistently towards 1, there is an higher dispersion of values in each bin when compared with featured articles. The height of each boxplot indicates the spread of values in each bin. Contrasting the median cosine similarity for both types of documents highlights the differences between the two datasets (see Figure 4).

4.2 Discretizing by Revision Date

In this approach, we consider the date information that is available in each revision made to a document. Based on this information, we can view the progression of similarity values as a function of *time* instead of *order*, as presented in the previous section. First, we discard short-lived documents by removing all articles with a total time span lower than 50 days. We also establish 25-quantiles for each article based on the complete temporal span of the article, from its inception to its current version. For instance, for an article spanning over 100 days, we first extract the revision made in the 4th day (1st bin), then the revision made in the 8th day (2nd bin), and so forth. If no revision was made on a specific day, we consider the most recent previous revision that was active on that day. The content found in each of the 25 days (bins) is compared to the article's latest revision, corresponding to the 100th day. Figure 5 represents a series of boxplots, each one summarizing the similarity values found between each bin and the current version of the document, for all featured articles.

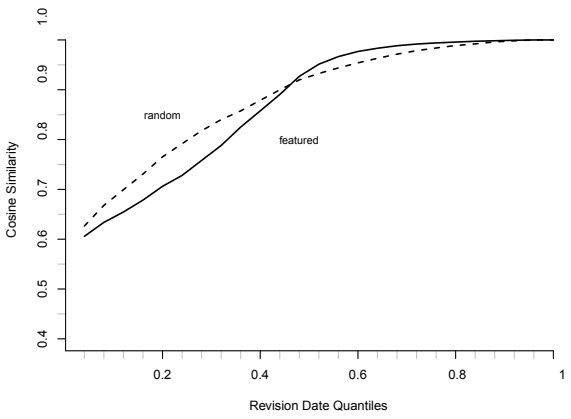

Figure 6: Median cosine similarity of featured and random articles, discretized by revision date.

The overall profile exhibits a more irregular evolution when compared with the projection based on revision order (contrast with Figure 2). In this case, the initial progression of similarity values is less delimited. The height of each boxplot shows a high dispersion in the values of each bin over a large part of the initial quantiles. A very fast convergence is noticeable close to the middle of the overall lifespan. As presented in Figure 6, the set of random articles shows a more regular progression in similarity values. Comparing Figures 4 and 6, we can see that a projection based on revision dates results in more visible differences in the set of featured articles.

To better understand the impact of being added to the *featured article* category, we determine the average bin that represents the moment when an article is associated to this category. When the horizontal bins are based on the revision order, the mean value is 23.2 and the median value is 24. This means that the large majority of revisions are made before the article is added to the *featured article* category. When bins are based on the revision date, the mean is 18.8 and the median is 20. The two vertical lines in Figures 2 and 5 represent the mean (solid line) and the median (dashed line) for each case. The lower value found in the second case indicates that the revisions made after being added to the *featured article* category are more dispersed through time. It is important to note that this information is not error-free. As mention in Wikipedia's documentation, if an article is vandalized and the category information is removed, the original dates for each category association are lost. Thus, it is likely that in reality these values are lower.

5. DISCUSSION AND CONCLUSIONS

In this paper we present an investigation about the internal dynamics of collaborative web documents. We analyze the complete lifespan of real web documents by comparing the progression of similarity values over two projection axes – revision order and revision time. Also, using the categories defined in Wikipedia, we contrast high quality document with regular documents. Based on this study we find that term frequency in encyclopedic documents – i.e. comprehensive and focused on a single topic – exhibits a rapid and steady progression towards the document's current version. The content in early versions quickly becomes very similar to the present version of the document. This contrasts with Adar et al's [1] work on the stable and dynamic content of popular documents on the web. While popular documents (e.g. portals, news sites) have a stable content core and a dynamic portion corresponding to the page sections that are regularly updated, in-depth focused articles don't have this structure and dynamics. The evolution of content is cumulative and centered on a principal theme.

Detailed knowledge about the internal dynamics of documents is important and has several applications, ranging from the definition of re-crawling and re-indexing strategies to the design and development of tools to manage and maintain documents. As future work we plan to investigate additional projection variables (e.g. document size) and applications related to the prediction of document change and quality.

6. ACKNOWLEDGEMENTS

Sérgio Nunes was financially supported by Fundação para a Ciência e a Tecnologia (FCT) and Fundo Social Europeu (FSE - III Quadro Comunitário de Apoio), under grant SFRH/BD/31043/2006.

7. REFERENCES

[1] E. Adar, J. Teevan, S. T. Dumais, and J. L. Elsas. The Web Changes Everything: Understanding the Dynamics of Web Content. In *WSDM'09: Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 282–291, New York, NY, USA, 2009. ACM.

[2] J. L. Elsas and S. T. Dumais. Leveraging Temporal Dynamics of Document Content in Relevance Ranking. In *WSDM'10: Proceedings of the Third ACM International Conference on Web Search and Data Mining*, pages 1–10, New York, NY, USA, February 2010. ACM, ACM.

[3] D. Fetterly, M. Manasse, M. Najork, and J. L. Wiener. A Large-Scale Study of the Evolution of Web Pages. *Softw. Pract. Exper.*, 34(2):213–237, February 2004.

[4] A. Ntoulas, J. Cho, and C. Olston. What's New on the Web?: The Evolution of the Web from a Search Engine Perspective. In *WWW'04: Proceedings of the 13th International Conference on World Wide Web*, pages 1–12, New York, NY, USA, 2004. ACM Press.

[5] C. Thomas and A. P. Sheth. Semantic Convergence of Wikipedia Articles. In *WI'07: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, pages 600–606, Washington, DC, USA, 2007. IEEE Computer Society.