Term Weighting based on Document Revision History

Sérgio Nunes, Cristina Ribeiro, and Gabriel David

INESC Porto, DEI, Faculdade de Engenharia, Universidade do Porto.

Rua Dr. Roberto Frias, s/n. 4200-465 Porto, Portugal.

Emails: {ssn, mcr, gtd}@fe.up.pt

Abstract

In real-world information retrieval systems, the underlying document collection is rarely stable or definite. This work is focused on the study of signals extracted from the content of documents at different points in time for the purpose of weighting individual terms in a document. The basic idea behind our proposals is that terms that have existed for a longer time in a document should have a greater weight. We propose four term weighting functions that use each document's history to estimate a current term score. To evaluate this thesis, we conduct three independent experiments using a collection of documents sampled from Wikipedia. In the first experiment we use data from Wikipedia to judge each set of terms. In a second experiment we use an external collection of tags from a popular social bookmarking service as a gold standard. In the third experiment, we crowdsource user judgments to collect feedback on term preference. Across all experiments results consistently support our thesis. We show that temporally aware measures, specifically the proposed revision term frequency and revision term frequency span, outperform a term weighting measure based on raw term frequency alone.

1 Introduction

In real-world information retrieval systems, the underlying document collection is rarely stable or definite. For instance, in personal systems, such as files or e-mails stored in a computer, documents are routinely added, removed or edited. Similarly, in enterprise and public environments, the existence of shared repositories of information is a standard scenario, resulting in active collections of documents that are continually updated. In this work we propose and investigate features, derived from the dynamic characteristics of collections, for weighting the importance of a term in a document. Term weighting is a core task in information retrieval settings with direct impact in many higher-level tasks, such as automatic summarization, keyword extraction, index construction or topic detection. It is our goal to evaluate the core task of term weighting for individual documents, without focusing on any particular application such as indexing or retrieval.

In a time-dependent collection, we can gather individual temporal clues using many different approaches (Nunes, 2007). For instance, we can use metadata obtained from the number of accesses over time to estimate the overall importance of documents. Alternatively, we can observe the individual changes made to documents over time and acquire indications about the relative importance of isolated terms. This work is focused on the study of content-based features over time, i.e. terms extracted from the content of documents at different points in time. The basic idea behind our approach is to give more weight to terms that have existed for a longer time in a document. For instance, it is our intuition that a term that has subsisted in a document since its first version should be valued higher than a term that was introduced only in the latest revision made. In other words, our hypothesis is that a term's prevalence over time is a good measure of importance. To evaluate this theory, we conduct several experiments using a collection of documents from Wikipedia — a unique public resource of reference documents collaboratively built by millions of anonymous users. One of the most distinctive features of Wikipedia is the fact that the full revision history associated with each article is kept and fully available via an application programming interface (API). We use this API to prepare a collection of documents and retrieve the corresponding historic versions for parsing. We evaluate the proposed measures using three independent methods. In the first approach, we use data from Wikipedia itself to judge each set of terms. In the second method, we use an external collection of tags from a popular social bookmarking service as a gold standard. Finally, with the third method, we use feedback gathered from users to evaluate and compare our proposals against classic measures.

2 Related Work

Term weighting is one of the key techniques in the field of Information Retrieval with direct application in a number of important retrieval tasks (e.g. automatic summarization, keyword extraction, indexing) (Singhal, 2001). The first published works on term weighting date back to the late 50s with Lunh's seminal work on the automatic production of abstracts (Luhn, 1958). In this work, Luhn proposes that "the frequency of word occurrence in an articles furnishes a useful measure of word significance". Luhn argues that the "justification of measuring word significance by use-frequency is based on the fact that a writer normally repeats certain words as he advances or varies his arguments and as he elaborates on an aspect of a subject". This term weighting measure was tested and experimentally evaluated in the production of automatic abstracts. Improved term weighting schemes were developed in the following years (Salton & Buckley, 1988). The Okapi weighting scheme (Robertson & Walker, 1994) is one of the most widely used in current retrieval systems. Document-based term weighting schemes such as these, contrast with collection-based term weighting schemes such as the inverse document frequency (idf) (Jones, 1972).

While raw term frequency alone is a crucial component for term weighting, other signals have been tested and evaluated as potential improvements to this measure. Such signals include measures that explore the document's structural information (Robertson, Zaragoza, & Taylor, 2004), term proximity (Keen, 1992), or term position (Troy & Zhang, 2007) among others. In this paper we use each document's history, a relatively unexplored signal, as a source of additional information for document term weighting. In the following paragraphs we review the existing literature on the use of temporal features for term weighting and discrimination.

In a recent work, Elsas and Dumais (2010) evaluate the relationship between document dynamics and relevance ranking. Using a collection of top ranked web documents, the authors establish a connection 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. Based on this finding, the authors propose two methods that improve document ranking by leveraging content change. In the first approach, a query-independent method, they find that favoring dynamic pages leads to performance improvements. In the second method, a query-dependent technique, it is shown that favoring a document's static content (i.e. content that prevails over time) also results in performance improvements. Although this work is not directly focused on term weighting it introduces a distinction between the terms in a document based on their temporal properties.

Efron (2010) directly addresses the problem of term weighting in a collection with the use of temporal cues. This work is, to the best of our knowledge, the first one to study the impact of time in term weighting. The author focuses on the behavior of terms while the collection changes over time as new documents are added. A new global query-independent term weighting measure is proposed and evaluated against idf. This work differs from ours since it is focused on changes occurring at the collection to propose a global term weight, while we address changes in individual documents to propose document-level measures.

More similar to our work, is the recently published paper by Aji, Wang, Agichtein, and Gabrilovich (2010) on the use of a document's "authorship process" as a source of information about term importance. The authors propose a new term weighting measure, named RHA (Revision History Analysis), which extends raw term frequency counts by incorporating the document revision history. The RHA measure combines three parts: a global model, a burst model, and the standard term frequency model. Both the global model and the burst model use a cumulative count of term frequencies across all previous revisions, modified using a decay factor. This factor is adjusted so that terms in older revisions have a higher value. In our work we use the same source of temporal evidence — each document's revision history — to propose several different term weighting measures. While the RHA measure mixes three components to deal with revision bursts, our approaches are simpler and treat all revisions as equal. Also worth noting is the fact that our measures are all parameter-free, thus they can be directly applied without any optimization step. Moreover, while we evaluate the quality of the weighted terms in three experiments, RHA is evaluated in the context of relevance ranking, as an extension to BM25 and to a language model.

3 Term Weighting and Document History

To incorporate the temporal dimension of documents in a scoring function, we consider that each document d is composed of a set of revisions defined as $R_d = \{r_1, r_2, \cdots, r_n\}$. The first version of a document is represented as r_1 and the latest as r_n. Additionally, the set of revisions of a document d containing the term t is given by $R_{t,d} = \{r : r \in$

R_d and $tf_{t,r} > 0\}$, where $tf_{t,r}$ represents the frequency of term t in revision r. Except where otherwise noted, we treat the words *version* and *revision* as synonyms, both representing a specific instance of a document at a given point in time. A document's individual revision is represented as a tuple $(ts, terms)$, where ts is a date corresponding to the instant when the revision was published, and $terms$ denotes the contents of the document at that moment. The content is modeled as a *bag-of-words* ordered by term frequency. Consider the Wikipedia article on 'Information retrieval' as an illustrative example. This article has more than 650 words in its latest version. A bag-of-words representation of its content, ordered by term frequency, would be as follows: $terms = \{\{\text{information}, 45\}, \{\text{retrieval}, 44\}, \{\text{documents}, 32\}, \{\text{relevant}, 17\}, \cdots\}$.

3.1 Revision Frequency

A weighting function incorporating a term's *revision frequency* (rf) is defined in Equation 1. Basically, a term's rf weight for a given document is defined as the ratio of the number of revisions containing that term to the document's total number of revisions. A term occurring in all versions of a document would have a rf score equal to 1. This measure ignores the frequency of terms at each revision, and only considers the presence or absence of the term. For instance, a term occurring multiple times at a given revision is weighted equally to a term appearing only once at that same revision. To incorporate a term's frequency at a given revision, we extend the previous formula and obtain a term's *revision term frequency* (rtf), as defined in Equation 2. In this case, we incorporate in the final score the *relative term frequency* (rel_tf) at each revision as defined by Equation 3.

In a nutshell, the rel_tf of a term in a document is defined as the ratio of the frequency of the term to the total number of terms in that document.

$$rf_{t,d} = \frac{|R_{t,d}|}{|R_d|} \qquad (1)$$

$$rtf_{t,d} = \frac{\sum_{r \in R_{t,d}} rel\_tf_{t,r}}{|R_d|} \qquad (2)$$

$$rel\_tf_{t,d} = \frac{tf_{t,d}}{\sum_{t\prime \in d} tf_{t\prime,d}} \qquad (3)$$

3.2 Revision Span

The previously defined term weighting measures view the revision history of a document as a set of evenly distributed document versions. However, the lifespan of each version varies widely, ranging from extremely short-lived versions (spanning over a few minutes) to long-lived versions that exist over many days. Taking this into account, we introduce the concept of *revision span* (rs), where the lifespan of each specific revision is taken into account in the weighting formula. This approach is defined in Equation 4, where the function $ts()$ is used to obtain a revision's date. In this case, the weight of a term in a document is defined as the ratio of the period when the term was in the document to the document's total lifespan. The numerator in Equation 4 gives the complete lifespan of a term in a document's revision history by adding the durations of all revisions containing the term. Finally, we extend this formula to also take into account the frequency of each term in each revision. This measure, named *revision term frequency span* (rtfs), is presented in Equation 5.

$$rs_{t,d} = \frac{\sum_{r_i \in R_{t,d}} (ts(r_{i+1}) - ts(r_i))}{ts(r_n) - ts(r_1)} \qquad (4)$$

$$rtfs_{t,d} = \frac{\sum_{r_i \in R_{t,d}} \left(rel\_tf_{t,r_i} \times (ts(r_{i+1}) - ts(r_i)) \right)}{ts(r_n) - ts(r_1)} \qquad (5)$$

3.3 Preliminary Comparison of Measures

In this section we introduce four term weighting functions that are based on a document's revision history. Two kinds of functions are presented: the first type of measures does not take into account the effective lifespan of each revision; in the second case, the lifespan of each revision is included in the weighting formula. In addition, we consider two approaches with respect to the frequency of a term at each revision. First, we only consider if a term is present or not in each revision, next we consider the relative term frequency at each revision. It is interesting to note that all proposed measures cumulatively weight terms over time treating each revision equally. This means that a top scoring term might not exist in the current version of a given article, a scenario completely impossible if term weighting is solely based on the current revision. We chose to also consider these terms because we have no warranties about the quality of the latest version (e.g. it could be a vandalized revision). Thus, we have made no assumption regarding this and decided to maintain the terms not appearing in the current version of a document.

We perform a first exploratory examination of these weighting functions and compare them with the classic *term frequency* measure (tf) by looking at a few illustrative examples, presented in Table 1. This table lists the 5 best scoring terms in Wikipedia articles obtained using each approach. We see that there are clear differences between each pair of methods, even when just the top 5 terms are considered.

Article	tf	rf	rtf	rs	rtfs
Information retrieval	information retrieval documents relevant precision	ir retrieval information science system	information retrieval documents ir text	ir acm science retrieval databases	information retrieval documents ir text
Research	research hypothesis scientific academic work	research information basic applied generally	research hypothesis basic academic scientific	information knowledge science applied research	research basic knowledge applied information
Data mining	data mining patterns analysis information	mining data large patterns analysis	data mining analysis information patterns	people correlations mining investment large	data mining analysis people information

Table 1: Results obtained with each method for different documents.

4 Experimental Evaluation

In this section, we present the methods designed to evaluate the proposed weighting measures. We adopt three independent approaches, the first based on Wikipedia data, the second based on a reference external collection and a third approach based on direct user feedback. We start with an analysis of the document collections and present some descriptive statistics. Then, we evaluate the impact of each scoring function on result diversity. Finally, in the last three sections, we document the evaluation experiments and discuss the corresponding results.

4.1 Document Collections

To evaluate the usefulness of the proposed measures, we use three independent sets of documents obtained from the English version of Wikipedia (http://en.wikipedia.org). The most important reason for choosing Wikipedia is the fact that the complete revision history for each article is kept and easily available via a public API. Additionally, Wikipedia is a very popular resource that includes many high quality documents, making it a popular object for research in subjects ranging from informatics to sociology. Finally, the fact that all content from Wikipedia is public guarantees that this study is reproducible by others.

We define three reference sets of documents for this research. The first set contains a random sample of Wikipedia featured articles, i.e. articles sampled from the 'Featured articles' category. A second set includes articles obtained via the 'Random article' feature available on Wikipedia. The third set is based on the most popular Wikipedia articles bookmarked at a well-known social bookmarking web site. This set was prepared using the Wiki10+ dataset released by Zubiaga (2009), which contains more than 20,000 unique Wikipedia articles, all of them with their corresponding social tags. Each set comprises a total of 100 distinct articles. A brief summary of the main properties of each set is presented in Table 2. The numbers included in the table represent the mean value for each attribute. The total number of words was calculated based on each article's current version. Comparing the different properties, we see a significant difference in the number of revisions between the random set and the other two sets. Interestingly, although articles in the social set have the highest number of revisions and age, they have fewer words than the articles in the featured set. This can be explained by the fact that featured articles need to meet certain criteria before being labeled as such. On the other hand, the social set includes articles that attract significant attention, which can explain the high number of revisions.

Set	**N**	**Revisions**	**Age (days)**	**Words (current)**
Featured	100	1199	2053	1199
Random	100	47	1199	140
Social	100	2415	2376	744

Table 2: Summary statistics for each set of documents.

4.2 Divergence in Scoring Functions

To observe the differences between the proposed measures, we computed the number of common terms in the rankings obtained with each pair of measures. The results for featured articles are outlined in Table 3. Although this table is symmetric, we have included all redundant values to facilitate reading. For each pair of scoring functions we determined the ratio of common items for a fixed number of top terms, specifically 10, 50 and 100. For instance, looking at this table, we can see that there are only 17% of items in common between the top 10 items ranked with tf and rs. We have highlighted the pairs with highest similarity. We can see that the use of term frequencies versus simple term existence is determinant. Also, the relatively low overall ratios suggest that the proposed measures introduce a noticeable number of new terms. Even with rtf, which has the highest overall similarity with tf, approximately 20% new terms are introduced.

	rf			**rtf**			**rs**			**rtfs**		
top	10	50	100	10	50	100	10	50	100	10	50	100
tf	26	36	41	**83**	**78**	**77**	17	30	36	70	63	63
rf		—		30	45	52	**59**	**75**	**78**	36	53	59
rtf	30	45	52		—		20	38	47	**79**	**75**	**75**
rs	**59**	**75**	**78**	20	38	47		—		26	50	59
rtfs	36	53	59	**79**	**75**	**75**	26	50	59		—	

Table 3: Percentage of common items between measures in featured articles.

4.3 Evaluation with Wikipedia Data

We can use Wikipedia itself to evaluate the quality of each set of terms. The idea is to use an article's lead as a summary of the body of the article. As stated in Wikipedia's Manual of Style (Wikipedia, n.d.) — "The lead should define the topic and summarize the body of the article with appropriate weight.". Given that featured articles are more likely to comply with Wikipedia rules, we assume that these articles have the best leads. Thus, we base this evaluation on the collection of featured articles. For each article in this set, we extract its lead (i.e. the first paragraph) and, for each approach, determine the number of terms found in it. We conduct this procedure for different numbers of top terms, as depicted in Figure 1. The *x-axis* represents the number of terms used and the *y-axis* the ratio of terms found in the article's lead. The numbers presented are the mean values over all 100 articles in the featured set.

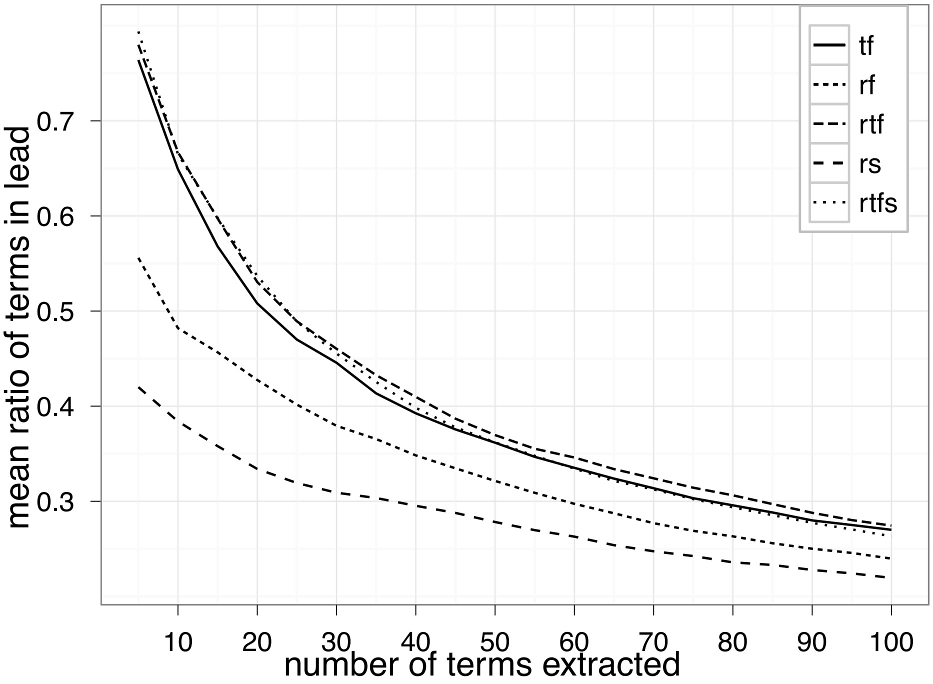


Figure 1: Mean ratio of terms found in articles' lead.

From this figure we can see that the measures with best performance are those based on the frequency of terms, as opposed to those based on the occurrence of terms. More important, we can see that both rtf and rtfs outperform the tf measure, when up to 50 terms are being tested. For more than 50 terms, the results obtained with rtfs decay slightly more rapidly than those obtained with tf. To evaluate the significance of these results, we use two sample paired t-tests for the rtf and rtfs measures with tf. Results are presented in Table 4, where each line represents a test using a specific number of top terms. From this table we can see that most results for rtf are significant, either at 95% or 99% — indicated with single or double asterisks respectively. For the rtfs measure, we only include the values where rtfs outperforms tf (up to 50 terms). Contrary to the rtf measure, the improvements obtained with rtfs are not significant (except for 20 terms). In summary, the evidence from this experiment shows that rtf is consistently better than tf for term extraction.

	rtf		**rtfs**	
terms	t(99)	p-value	t(99)	p-value
10	1.767	0.040*	1.122	0.132
20	2.839	0.003**	2.506	0.007**
30	1.862	0.033*	0.995	0.161
40	2.772	0.003**	0.723	0.236
50	1.531	0.064	0.055	0.478
60	2.150	0.017*	—	—
70	2.400	0.009**	—	—
80	2.597	0.005**	—	—
90	2.118	0.018*	—	—
100	1.311	0.096*	—	—

Table 4: Paired t-test results for rtf and rtfs versus tf using articles' leads.

4.4 Evaluation with Social Annotations

Wikipedia articles are very popular among Internet users. A significant number of articles is shared by users, either by e-mail, blog posting or social bookmarking. This observation is supported by a simple analysis of the Wiki10+ dataset released by Zubiaga (2009). This dataset was prepared in April 2009 and includes all articles from the English version of Wikipedia that were bookmarked in Delicious (http://delicious.com) by at least 10 users. Delicious, currently a Yahoo! property, was a pioneer service in the area of social bookmarking and is still considered one of the references in this area. The Wiki10+ dataset contains 20,764 unique URLs and, for each URL, all corresponding Delicious tags. A simple analysis based on the histogram shown in Figure 2 reveals that the dataset only includes up to 30 tags for each bookmark. This can be explained by the fact that Delicious only displays the 30 most popular tags for each bookmark and offers no other way of obtaining the complete set of tags. Table 5 presents the 10 most popular tags found in this dataset for the three articles considered in Section 3.3. It is worth noting that some of the tags used are simple graphical variations of each other (e.g. data-mining and data_mining). We make no effort to consolidate or correct these instances.

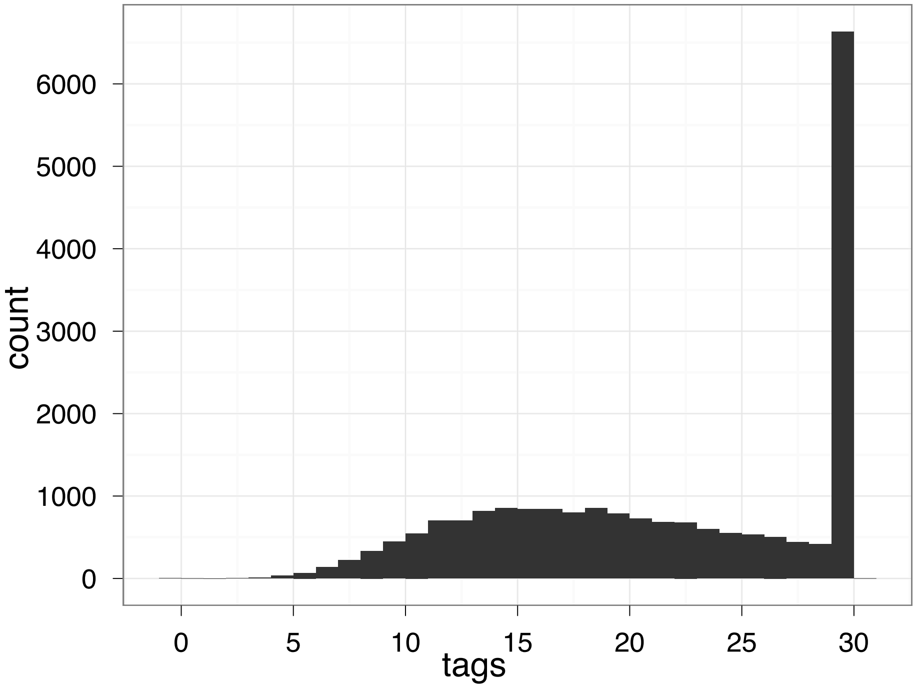


Figure 2: Distribution of bookmarks by number of tags.

Article	Top Delicious Tags	
Information retrieval	search	reference
	ir	informationretrieval
	information-retrieval	retrieval
	information	research
	wikipedia	recall
Research	research	dissertation
	wikipedia	terminology
	science	overview
	definition	researching
	info	science_technology
Data mining	datamining	database
	wikipedia	programming
	data	statistics
	mining	data_mining
	reference	data-mining

Table 5: 10 most popular tags on Delicious for different articles.

To evaluate each term weighting approach using the Delicious external reference set, we measure the number of common items pairwise. First, we select the 100 bookmarks in this dataset with the highest number of users — i.e. those that were bookmarked by more users. Then, we compare the tags available for each bookmark with the terms extracted using each method. Figure 3 summarizes the results obtained, presenting the percentage of common items found for different numbers of top terms. We can see that both rtf and rtfs have a higher number of terms in common with the Delicious set. The superiority over tf is consistent across all number of terms considered. Again, the worst performing measure is rs. Given that, for each term extraction method, we have weights associated with each term, we can use this information to make a more precise comparison with each tag's weight found in Delicious. Thus, for each one of the 100 articles, we produce a weighted term vector using all tags found on the Wiki10+ dataset. Then, for each term extraction method and for each article, we also create term vectors considering a different number of top terms. Specifically, we build four vectors for each article and method, one including all terms and the others considering only the top 10, 50 and 100 terms. Finally, we calculate the cosine similarity between the reference vector based on Delicious data and each of the five vectors. The results, averaged over all articles, are presented in Table 6.

The rtf method outperforms all other methods, including the reference tf measure. We use a two sample paired t-test to evaluate the significance of rtf's performance over tf. We find that rtf's better performance when using all terms is significant at 99% (t(99)=3.78, p=0.0001), and significant at 95% when restricting the vector to the top 10 terms (t(99)=2.24, p=0.014) and the top 100 terms (t(99)=1.96, p=0.026). Again, using a different experimental setup, we see that a time-aware measure exhibits better results than an approach that discards historical information.

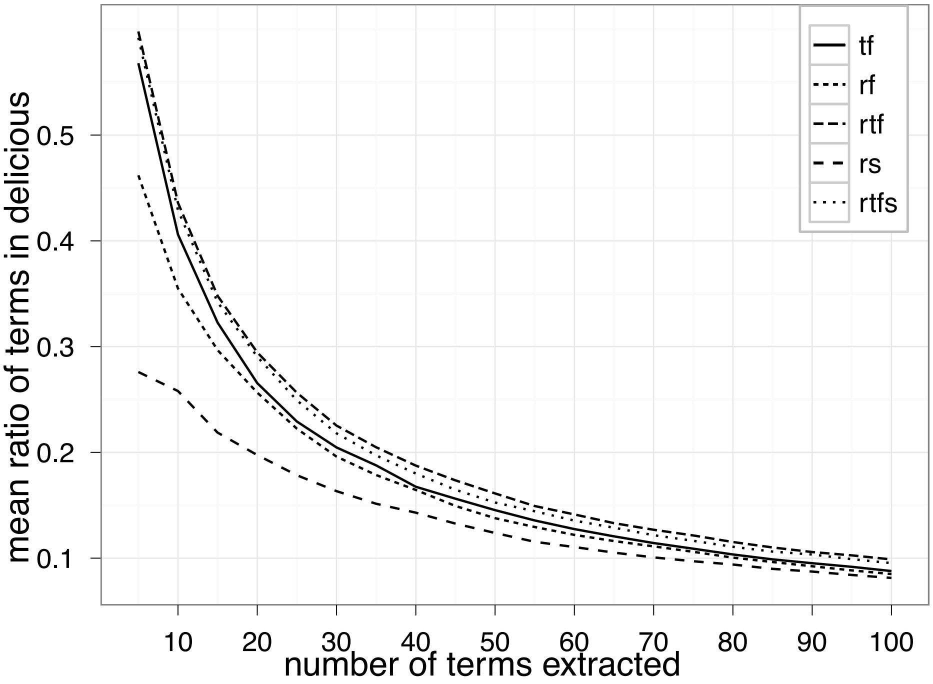


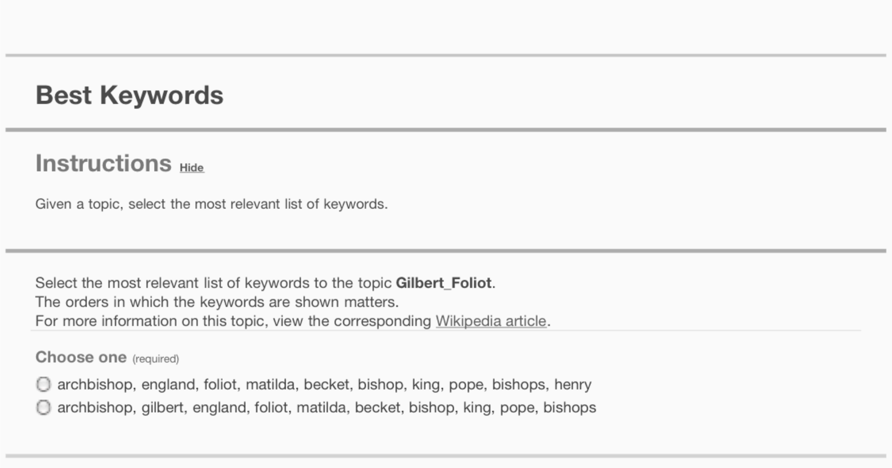
Figure 3: Mean ratio of terms found in top Delicious tags.

	top 10	top 50	top 100	all
tf	.441	.428	.422	.408
rf	.273	.199	.168	.127
rtf	**.459**	**.437**	**.436**	**.436**
rs	.185	.183	.164	.130
rtfs	.444	.419	.419	.419

Table 6: Cosine similarity between Delicious tags and each method’s terms.

4.5 Evaluation with User Feedback

The previous evaluation methodologies are based on indirect measures, i.e. no direct user feedback is collected. In this section we describe an evaluation experiment designed to obtain direct user judgments. Basically, for each article in the evaluation set, we present two alternative lists of terms and ask the user to choose the most relevant to the article. We do some basic stop word removal and then extract an ordered list of 10 terms using each algorithm. We use the crowdsourcing (Howe, 2006) service CrowdFlower (http://crowdflower.com) to design this experiment and collect user feedback. CrowdFlower is a service that redirects user-designed tasks to 'labor-on-demand' marketplaces, such as Amazon's Mechanical Turk. These tasks, known as Human Intelligence Tasks (HITs), are distributed across Internet users (i.e. workers) that execute them in exchange of monetary payment. Given the lack of direct supervision, the execution of individual tasks offers no assurance in relation to quality control. It is well known that task design and indirect control mechanisms, such as qualification tests, are paramount when crowdsourcing jobs (Kittur, Chi, & Suh, 2008). To improve the quality of our results we try to eliminate low-value work by using two different strategies: request multiple judgments for each task and define some tasks as ground truth.



Best Keywords

Instructions

Given a topic, select the most relevant list of keywords.

Select the most relevant list of keywords to the topic **Gilbert_Foliot**.
The orders in which the keywords are shown matters.
For more information on this topic, view the corresponding Wikipedia article.

Choose one (required)

○ archbishop, england, foliot, matilda, becket, bishop, king, pope, bishops, henry
○ archbishop, gilbert, england, foliot, matilda, becket, bishop, king, pope, bishops

Figure 4: Interface design for evaluation task in CrowdFlower.

A screen capture of the interface presented to workers is pictured in Figure 4. For each individual assessment task, we require a minimum of 5 independent judgments. Using this information, we only consider valid answers those where the most voted option wins by at least $^2/_3$ of the votes. Additionally, we define 10% of the tasks of a given pairwise comparison as ground truth, known as *gold* in CrowdFlower (https://crowdflower.com/docs/gold). Setting gold tasks can substantially improve the quality of the answers. CrowdFlower's proprietary algorithms use this information, together with worker's historical record, to automatically accept or reject submissions. Given that we are conducting subjective tasks, there is no correct answer to use as ground truth. Thus, we create artificial tasks that we use as ground truth. To produce these tasks we simply replace one of the term lists with a list of keywords obtained from an unrelated article. For instance, when evaluating an article on the NeXT computer system, the user is presented with a list containing correct terms (e.g. nextstep, computers, jobs) and another with off topic terms obtained from a completely unrelated article (e.g. lancelot, merlin, excalibur). We mark the first option as the correct choice and define this task as gold in CrowdFlower's interface.

Considering the monetary costs associated with this experiment, we select a subset of 50 articles from the original collection of 100 articles. After running the experiments, we simply count the number of wins for option 1 versus option 2. In Table 7, we present the confidence intervals at 95% for the true proportion of wins of each new measure over the term frequency baseline. These intervals are calculated approximating the binomial distribution to the normal distribution. For instance, when considering featured articles, we are 95% confident that the interval 51%—82% contains the true proportion of wins of rtf over tf. As the lower confidence limit of this interval is higher than 50%, we can state that the rtf measure is preferred over the tf one, outperforming it. The same happens with the rtfs measure when compared with tf.

Overall, the quality of the proposed measures is clearer when considering the set of featured articles. In addition, we see that the rtf measure performs worst than tf for the set of social articles. We think that this can be explained by the fact that the articles in the social set are more vulnerable to vandalism and subsequent reverts. Thus, a measure that ignores the duration of the revisions (like rtf) is likely to be affected by this. We can see from Table 2 that the articles in the social set have a much higher number of revisions, despite a similar age and a significantly lower current number of words. To conclude, we can say that these results are clear and consistent with those reported in the previous experiments. Again, we see that the use of document history in term weighting algorithms consistently improves the results.

	rf	rtf	rs	rtfs
Featured	(0.105, 0.335)	(0.513, 0.821)	(0.089, 0.311)	(0.538, 0.809)
Social	(0.105, 0.335)	(0.317, 0.599)	(0.138, 0.382)	(0.433, 0.710)

Table 7: Confidence intervals at 95% for users' preferences for each method versus term frequency.

5 Conclusions

In this work we have studied the influence of document history in term weighting. We define and extensively evaluate four new measures for document term weighting. All the proposed measures explore the document's revision history as an additional signal to improve term discrimination. Based on different evaluation experiments we show that document history is a useful source of information to improve document term weighting. We demonstrate that temporally aware measures, specifically the proposed *revision term frequency* and *revision term frequency span*, outperform the tf measure. Although we have used Wikipedia, and the full revision history of its articles as a document collection, this work can be easily adapted to other contexts. Consider the case of web search. Given that web search engines periodically crawl the web, they have access to historical information about web documents. This information can be used without difficulty to incorporate time-dependent signals on term weighting functions.

It is worth noting that traditional measures like term frequency are based on a single version of a document (i.e. the current version), thus directly dependent on the latest updates. On the contrary, the proposed time-dependent measures are based on multiple versions of the same document. This results in more robust weighting measures, which are less vulnerable to sporadic changes. This is a valuable quality in the context of shared or public repositories because of the higher resistance to SPAM or other malicious modifications. Nonetheless, this robustness can be seen as a drawback when dealing with naturally fast changing documents like homepages that are continually updated with the latest information.

Finally, we would like to highlight the full reproducibility of this work. All data, except for the human assessments, is public and freely available.

6 Acknowledgments

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. We thank the anonymous reviewers, whose comments have contributed to important improvements to the final version of the paper.

References

Aji, A., Wang, Y., Agichtein, E., & Gabrilovich, E. (2010). Using the past to score the present: extending term weighting models through revision history analysis. In *Proceedings of the 19th ACM International Conference on Information and*

Knowledge Management (ACM CIKM'10) (pp. 629–638). New York, NY, USA: ACM.

Efron, M. (2010). Linear time series models for term weighting in information retrieval. *Journal of the American Society for Information Science and Technology*, 61(7), 1299–1312.

Elsas, J. L., & Dumais, S. T. (2010). Leveraging temporal dynamics of document content in relevance ranking. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (ACM WSDM'10)* (pp. 1–10). New York, NY, USA: ACM.

Howe, J. (2006, June). The rise of crowdsourcing. *Wired magazine*, 14(6).

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1), 11–21.

Keen, E. M. (1992). Term position ranking: Some new test results. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR'92)*. New York, NY, USA: ACM.

Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk. In *Proceeding of the 26th Annual SIGCHI Conference on Human Factors in Computing Systems (ACM CHI'08)* (pp. 453–456). New York, NY, USA: ACM.

Luhn, H. P. (1958, April). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2), 159–165.

Nunes, S. (2007). Exploring temporal evidence in web information retrieval. In *BCS IRSG Symposium Future Directions in Information Access (FDIA'07)* (pp. 44–50). Cambridge, England: BCS IRSG.

Robertson, S. E., & Walker, S. (1994). Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR'94)* (pp. 232–241). New York, NY, USA: Springer-Verlag New York, Inc.

Robertson, S. E., Zaragoza, H., & Taylor, M. (2004). Simple BM25 extension to multiple weighted fields. In *Proceedings of the 13th ACM International Conference on Information and Knowledge Management (ACM CIKM'04)* (pp. 42–49). New York, NY, USA: ACM.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5), 513–523.

Singhal, A. (2001). Modern information retrieval: A brief overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 24(4), 35–42.

Troy, A. D., & Zhang, G. Q. (2007). Enhancing relevance scoring with chronological term rank. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR'07)* (pp. 599–606). New York, NY, USA: ACM.

Wikipedia: Manual of style (n.d.). In *Wikipedia*. Retrieved December 6, 2010, from http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style

Zubiaga, A. (2009, August 26-28). Enhancing navigation on Wikipedia with social tags. In *Wikimania 2009*, Buenos Aires, Argentina.