Comparing and Evaluating Information Retrieval Algorithms for News Recommendation

Toine Bogers
ILK, Tilburg University
P.O. Box 90153, 5000 LE
Tilburg, The Netherlands
A.M.Bogers@uvt.nl

Antal van den Bosch
ILK, Tilburg University
P.O. Box 90153, 5000 LE
Tilburg, The Netherlands
Antal.vdnBosch@uvt.nl

ABSTRACT

In this paper, we argue that the performance of content-based news recommender systems has been hampered by using relatively old and simple matching algorithms. Using more current probabilistic retrieval algorithms results in significant performance boosts. We test our ideas on a test collection that we have made publicly available. We perform both binary and graded evaluation of our algorithms and argue for the need for more graded evaluation of content-based recommender systems.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*performance evaluation*

General Terms

Experimentation, measurement, performance

Keywords

Recommender systems, information retrieval, evaluation, language modeling, probabilistic IR, news recommendation

1. INTRODUCTION

During the first quarter of 2007 more than 59 million people (37.6% of all active Internet users) visited over 2000 available newspaper Web sites in the US alone [6, 17]. This number has been steadily increasing over the past decade and shows the growing appeal of reading news online. Many newspapers post at least a subset of their hard-copy articles online and, depending on online subscription schemes, sometimes all of them. One of the advantages of adding articles to a newspaper's website is that the online versions can be augmented with extra information, such as links to recommended related articles. Finding these related articles is currently usually a time-consuming job with editors manually searching for related articles.

Recommending related online content from the same website or domain is not only a useful functionality for newspaper websites.

For instance, consumer portals of insurance or travel agents and educational websites could also use this technology to recommend related products automatically. In this paper we describe ongoing work in automatically recommending related news articles. We view article recommendation as a specific type of information retrieval (IR) task. The recommender algorithms described in this paper are performed independently of the user.

In earlier work on news recommendation and recommender systems in general, two main approaches can be distinguished: collaborative filtering and content-based filtering. Research into recommender systems has predominantly focused on collaborative filtering techniques, where user-item preference data (such as ratings or purchase data) from a group of users are used to make individual predictions. Collaborative filtering is useful in situations where content analysis is hard, e.g. in multimedia recommendation. In our news recommendation scenario, however, we have no user-item preference data. Instead, we focus on content-based recommendation, by approaching the problem from an IR perspective in which recommendations are based on matching textual content. In our experiments we examine three areas that we believe content-based recommendation could benefit from the most:

- Most content-based recommender systems use **relatively old or simple retrieval models** such as simple keyword matching or the vector space model with basic tf·idf weighting. These are known to have been outperformed by newer approaches such as language modeling [14].

- There have been no systematic experiments with the **representation length** of the news articles in news recommendation. Is it better to only use the article titles, or should matching be performed at the full-text level?

- System-driven **evaluation** of different content-based news recommenders has been **difficult to compare** since all approaches use different data sets. In addition, evaluation efforts have been divided between binary and graded relevance evaluation, but rarely together which makes them even harder to compare objectively.

We will discuss each of these points in more detail in sections 3–5. We start by describing the creation of the test collection we used to evaluate our recommender system. Sections 6 and 7 contain the experimental results and conclusions, respectively.

2. CONSTRUCTING A RECOMMENDER TEST COLLECTION

Content-based recommendation has been approached in various ways. The approach taken tends to determine the way the test collection is built. For instance, Mooney et al. looked at the problem

from a text categorization perspective in their book recommender system [13], so they gathered a labeled data set. Document clustering has also been used for recommendation, as related documents are likely to be found in the same cluster [3]. Clustering can be regarded as a particular instance of similarity-based IR methods, and requires an unlabeled document set as training material and query-document set pairs for evaluation [16]. Collaborative filtering has also been used often in news recommendation [12] and has proven successful on a very large scale [8]. In our scenario we do not have user-preference data, so our situation is likely to benefit most from a topic-centric IR approach.

Even though we approach recommendation as an IR problem and as such use document similarity to find related documents, we cannot use regular test collections. IR test collections based on news articles have been used in the past, e.g. in the Ad Hoc tracks of TREC 1-5 [20], containing short queries, such as "What progress has been made in fuel cell technology?", coupled with sets of related articles. In contrast, our recommendation task requires the *full* documents to be labeled as related or unrelated to other full documents. We therefore chose to create our own collection using the Reuters RCV1 collection. This collection contains 806,791 news articles published between August 20, 1996 and August 19, 1997.

Our approach to creating the test collection was different from [16] in that we did not use a combination of metadata and relevance feedback to create the queries and their corresponding relevance judgments. Instead, we used complete documents as our queries and aimed to find the related documents for those *focus articles*. We specifically focused on relatedness instead of relevance: the two concepts are likely to be correlated, but we do not assume them to be identical.

Based on the TREC pooling approach [20], we used three different IR algorithms to create a pool of potentially relevant documents for 50 query articles that we randomly selected from the ∼807K articles. For each query article, the document rankings from the three IR algorithms were merged and the top 100 results were selected to be judged. In accordance with the TREC pooling procedure, the non-judged documents were considered irrelevant. We invited colleagues to participate in judging. The only difference with the TREC pooling approach is that here each document was judged by only one person. After a short briefing on our news recommendation scenario, participants were asked to judge the relatedness between each of the 100 recommended articles and the focus artice on a 5 point scale. A score of 0 meant the articles were not related; 1 – slightly related; 2 – fairly related; 3 – very related; and 4 – highly related. We did not ask participants to take temporal aspects into account when judging the relatedness of two articles.

3. RECOMMENDER ALGORITHMS

An analysis of earlier work on content-based recommendation from an IR perspective reveals that relatively old or simple retrieval models have been used, such as simple keyword matching [6]. More than half of the news recommenders Montaner et al. compare in their 2003 paper use some form of the vector space model with tf·idf weighting [12]. More recently, Ha also reports using tf·idf [10]. More advanced algorithms such as probabilistic IR models or language modeling have hardly been used, with [11] as a notable exception. Lavrenko et al. used language modeling to predict which stories are likely to influence the financial markets.

In our experiments we have compared two retrieval models that have been underutilized in content-based recommendation to **tf·idf**. The first algorithm is the Okapi retrieval function (**okapi**), which has been proposed as an effective retrieval formula that represents

the classical probabilistic retrieval model [15]. The second is the language modeling framework (**LM**) as introduced by [14] which builds a probabilistic language model from each document d, and ranks documents on query likelihood: the probability of the model generating the query. Preliminary experiments suggested using the Kullback-Leibler divergence metric and Jelinek-Mercer smoothing in our experiments. The baseline system with tf·idf term weighting used in our experiments uses document similarity to find relevant documents and takes into account term frequency, its distribution across the collection, and the document length in calculating the weights [15]. We are not aware of any work comparing these three algorithms in the specific context of (news) recommender systems.

4. ARTICLE LENGTH

To the authors' knowledge there have been no systematic experiments with the article representation length in news recommender systems. Some systems use the entire article [6] while other systems only consider the article title [19]. However, document length has played a notable part in IR experiments over the years. Singhal et al. showed the weakness of the cosine function in tf·idf-based retrieval for very long documents [18]. Bennet et al. report similar influences of document length on text categorization in [1]. Callan's passage retrieval experiments were also motivated by the idea that document length influences the retrieval process [4].

It stands to reason that article length also plays a role in the recommendation process. News articles are organized in so-called *inverted pyramid style*, meaning that the most important information is at the beginning [9]. Cutting off the tail of an article would imply removing relatively less crucial information, and the article length reduction might speed up the recommendation process as well.

We performed a series of systematic experiments on the collection varying the length of the articles. In theory, three combinations are possible here. One option is to vary the length of the collection articles, but use the full focus article text. The other way around is also an option. However, this is much less practical: the indexing cost of one query article is marginal compared to indexing 800,000+ articles, so using the full focus article text is the more efficient option (which we refer to as *F-fix*). A third option is constraining the length of both the focus article and the collection articles simultaneously at the same threshold (*F-var*). We experimented with both the F-fix and the F-var options and the smallest variant in both runs contained only the document title (T+S00), the second variant the document title and the first sentence (T+S01), the third variation the title and the first 2 sentences of the articles (T+S02), and so on up to the title and the first 10 sentences (T+S10). Using larger slices of text would have resulted in too many articles not having enough text, resulting in an unfair comparison. Section 6.2 lists the results of these experiments.

5. EVALUATION

The lack of standardized content-based recommender test collections has made system-driven evaluation and comparison of different approaches difficult. Introducing yet another collection in section 2 does nothing to alleviate this problem if it is not made public. Therefore, we have made our collection publicly available in the hope that it can then be used for other (user-independent) news recommender experiments.[1]

Another obstacle to a fair comparison is that most evaluation efforts in the past have been divided between binary and graded relevance evaluation and rarely both of them at the same time, making

---

it harder to compare systems objectively. Binary evaluation is often mentioned in the literature [2, 13]. We suspect this is because (especially implicit) user feedback is harder to translate to graded relevance judgments than to binary labels. In the binary evaluation of the experiments described in this paper, we converted the document relatedness ratings to a binary relatedness scale. Preliminary experiments showed that a good threshold was to regard articles rated 3 or 4 as related, and lower scores as unrelated. This resulted in an average of 31.7 related documents per focus article.

We used the mean uninterpolated average precision (MAP) measure to perform the binary relevance evaluation. MAP is the mean of the precision scores obtained after each relevant document is retrieved, using zero as the precision for relevant documents that are not retrieved.

Graded evaluation is used by [6] and [13] among others. The attractiveness of performing graded evaluation lies in the fact that relevance (or relatedness) is not simply a binary concept: recommendation relevance occurs in different gradations. Because we collected our judgments on a 5-point graded scale, we also performed graded evaluation by correlating the gold standard ranking with the system's output. Popular measures of rank correlation are Spearman's rank correlation and Kendall's tau. We used Kendall's tau because the distribution of this statistic has slightly better statistical properties [7]. However, in almost all situations the values of Spearman's rank correlation and Kendall's tau are very close and will lead to the same conclusions. We compare MAP and Kendall's tau in section 6.3.

6. RESULTS

6.1 Recommender algorithms

We compared the three recommender algorithms by performing basic document retrieval using each of the 50 focus articles as queries. The only metadata we included in the article representations were the title and the body of the article. We experimented with adding other metadata (e.g. location, author, and topic codes) to the article representations, but both weighted and unweighted these additions did not produce significant performance gains.

Our baseline **tf·idf** system achieved a MAP of 0.6136, but **okapi** had a significantly higher score of 0.7016 ($p < 0.001$), an improvement of 14.33%. **LM** also improved significantly over **tf·idf** ($p < 0.013$) with a score of 0.6973, a performance gain of 13.63%. This is consistent with the reported gains in the literature [14]. The difference between **okapi** and **LM**, however, was not significant ($p = 0.81$). This suggests that significant performance gains can be made in content-based recommendation simply by switching retrieval models.

Another factor to perhaps take into consideration here is the execution time, as recommendations should be generated as quickly as possible: **LM** was on average 5.5 times faster than **okapi**.

6.2 Article length

Table 1 shows the results of the experiments with article length. Using the title and the first couple of sentences provides the biggest jump in performance compared to using only the title, but performance keeps increasing with the amount of information used. The best performances can be observed when using all of the focus and collection article text. This suggests that even though news articles tend to be written in inverted pyramid style, this does not necessarily mean that information can be thrown away safely for recommendation purposes.

In addition, the increase in performance does seem to level off

most for **tf·idf**, however, and even decreases in the F-fix run, although the difference is not significant. This is in line with Singhal's finding of **tf·idf** having problems with longer documents [18].

As for comparing the F-fix and F-var runs, always using the complete focus article as opposed to reducing it in size along with the collection articles tends to mildly improve performance for all the different combinations of collection article size. However, these improvements were only significant in the **okapi** case ($p < 0.0005$).

Table 1: MAP scores for the article size experiments, both the F-fix and F-var runs as described in section 4. Best scores are printed in bold.

	LM		okapi		tf·idf	
	F-fix	F-var	F-fix	F-var	F-fix	F-var
T+S00	0.1598	0.2825	0.2680	0.2681	0.2122	**0.3069**
T+S01	0.4599	0.4114	**0.4745**	0.4097	0.3900	0.4401
T+S02	**0.5076**	0.4512	0.5017	0.4450	0.4797	0.4780
T+S03	0.5367	0.4779	**0.5408**	0.4714	0.5093	0.5032
T+S04	0.5785	0.5198	**0.5805**	0.5177	0.5580	0.5254
T+S05	**0.6032**	0.5508	0.5978	0.5506	0.5849	0.5584
T+S06	**0.6225**	0.5940	0.6144	0.5819	0.6011	0.5894
T+S07	**0.6369**	0.6170	0.6349	0.6073	0.6171	0.6019
T+S08	**0.6426**	0.6223	0.6396	0.6201	0.6180	0.6059
T+S09	0.6505	0.6325	**0.6545**	0.6364	0.6219	0.6037
T+S10	0.6665	0.6362	**0.6683**	0.6409	0.6342	0.6098
all text	0.6973		**0.7016**		0.6136	

6.3 Binary vs. graded relevance

Table 2 shows the MAP scores for the different algorithms for the best-performing F-fix runs. Kendall's tau also increases with article representation length, albeit more steadily, for both **okapi** and **LM**. However, there is not a very clear increase in Kendall's tau, even though MAP scores increase more clearly. Table 2 clearly shows that using all text (significantly) decreases the performance of **tf·idf**: both the MAP and Kendall's tau scores decrease when using all of the text compared to using just a part of it.

Table 2: MAP scores compared to Kendall's for the three algorithms in the F-fix run. Best scores for each column are printed in bold.

	LM		okapi		tf·idf	
	MAP	K's τ	MAP	K's τ	MAP	K's τ
T+S00	0.1598	0.1431	0.2680	0.2467	0.2122	0.2212
T+S01	0.4599	0.2576	0.4745	0.2719	0.3900	0.2510
T+S02	0.5076	0.2253	0.5017	0.2420	0.4797	0.2293
T+S03	0.5367	0.2449	0.5408	0.2656	0.5093	0.2497
T+S04	0.5785	0.2483	0.5805	0.2749	0.5580	0.2573
T+S05	0.6032	0.2668	0.5978	0.2856	0.5849	0.2560
T+S06	0.6225	0.2717	0.6144	0.2833	0.6011	**0.2689**
T+S07	0.6369	0.2758	0.6349	0.2802	0.6171	0.2656
T+S08	0.6426	0.2779	0.6396	0.2882	0.6180	0.2622
T+S09	0.6505	0.2790	0.6545	0.2813	0.6219	0.2680
T+S10	0.6665	0.2822	0.6683	0.2871	**0.6342**	0.2630
all text	**0.6973**	**0.2932**	**0.7016**	**0.2968**	0.6136	0.2293

We have also examined the relation between MAP and Kendall's tau. Calculated over all queries and systems the correlation between the two measures is r(298) = .4865 ($p < 2.2 \cdot 10^{-16}$). This means the ratings are moderately correlated. Therefore, binary evaluation need not always lead to the same conclusions when evaluating a system compared to a graded evaluation.

7. CONCLUSIONS & DISCUSSION

In this paper we have presented work on comparing and evaluating news recommendation systems. We identified and examined three elements notably absent from the news recommendation literature. First, we showed that significant performance gains can be made by choosing more advanced, probabilistic retrieval algorithms such as language modeling and Okapi over the popular, but relatively old tf·idf weighting model.

We also examined the role that representation length can play in news recommendation, finding that for the probabilistic methods, using more of the article offers the best performance. This seems to suggest that relying on an linear style of writing alone is not good enough: using less text is detrimental to performance. We also showed that, in contrast to the above findings, the popular tf·idf model actually suffers from using all available text. It does, however, work better when there is little information to work with. We also showed that, in contrast, the popular tf·idf model actually suffers from using all available text. It does, however, work better when there is little information (i.e. only the title) to work with. We expect these findings to translate well into personalized content-based recommendation.

We evaluated our experiments using both binary and graded evaluation and both evaluation approaches turned out to be just moderately correlated. We argued that graded evaluation is intrinsically better, so we suggest that future evaluation of content-based recommender systems should at least be evaluated using graded measures. Finally, we also made available the test collection used in our experiments to facilitate experimentation on at least one fixed collection.

8. FUTURE WORK

In extending the current study we plan to examine the relationship between binary and graded evaluation further by looking at other measures such as Discounted Cumulated Gain. We would also like to experiment with generating personalized article recommendations. One possible way of doing this is simulating the browsing behaviour of an online reader and constructing personal recommendation profiles based on this behaviour. Other possible directions for future work are finding other collections with more metadata to exploit and combining our content-based approach with a collaborative approach to recommendation. Finally, investigating the influence of other performance indicators (such as article length) on the recommendation process might also be a worthwhile direction for future research.

9. ACKNOWLEDGMENTS

10. REFERENCES

[1] P. N. Bennett, S. T. Dumais, and E. Horvitz. The Combination of Text Classifiers Using Reliability Indicators. *Information Retrieval*, 8(1):67–100, 2005.

[2] D. Billsus and M. J. Pazzani. A personal news agent that talks, learns and explains. In O. Etzioni, J. P. Müller, and J. M. Bradshaw, editors, *Proceedings of Agents '99*, pages 268–275, New York, NY, 1999. ACM Press.

[3] T. Brants and R. Stolle. Finding similar documents in document collections. In *Proceedings of LREC-2002, Workshop on Using Semantics for Information Retrieval and Filtering*, 2002.

[4] J. P. Callan. Passage-level evidence in document retrieval. In W. B. Croft and C. van Rijsbergen, editors, *Proceedings of SIGIR '94*, pages 302–310, New York, NY, USA, July 1994. Springer Verlag.

[5] J.-H. Chiang and Y.-C. Chen. An intelligent news recommender agent for filtering and categorizing large volumes of text corpus. *International Journal of Intelligent Systems*, 19:201–216, 2004.

[6] M. Claypool, A. Gokhale, T. Miranda, P. Murnikov, D. Netes, and M. Sartin. Combining content-based and collaborative filters in an online newspaper. In *Proceedings of the SIGIR Workshop on Recommender Systems*, August 1999.

[7] W. J. Conover. *Practical Non-Parametric Statistics*. John Wiley and Sons, New York, NY, second edition, 1980.

[8] A. Das, M. Datar, A. Garg, and S. Rajaram. Google News Personalization: Scalable Online Collaborative Filtering. In *Proceedings of WWW '07*, pages 271–280, 2007.

[9] W. Fox. *Writing the News: A Guide for Print Journalists*. Iowa State University Press, second edition, 2001.

[10] S. H. Ha. Digital Content Recommender on the Internet. *IEEE Intelligent Systems*, 21(2):70–77, 2006.

[11] V. Lavrenko, M. Schmill, D. Lawrie, P. Ogilvie, D. Jensen, and J. Allan. Language models for financial news recommendation. In *Proceedings of CIKM 2000*, pages 389–396, New York, NY, 2000. ACM Press.

[12] M. Montaner, B. López, and J. L. de la Rosa. A Taxonomy of Recommender Agents on the Internet. *Artificial Intelligence Review*, 19(4):285–330, 2003.

[13] R. J. Mooney and L. Roy. Content-based book recommending using learning for text categorization. In *Proceedings of DL '00*, pages 195–204, New York, NY, 2000. ACM Press.

[14] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proceedings of SIGIR '99*, pages 275–281, New York, NY, 1998. ACM Press.

[15] S. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of SIGIR '94*, pages 232–241, New York, NY, 1994. ACM Press.

[16] M. Sanderson. Reuters test collection. Technical report, Glasgow University Comp. Sc. Department, June 1994.

[17] J. Sigmund. Online Newspaper Audience Sets Records in First Quarter, April 2007. Last visited: June 2007, http://www.naa.org.

[18] A. Singhal, G. Salton, M. Mitra, and C. Buckley. Document length normalization. *Information Processing and Management*, 32(5):619–633, 1996.

[19] N. Tintarev and J. Masthoff. Similarity for news recommender systems. In *Proceedings of the AH'06 Workshop on Recommender Systems and Intelligent User Interfaces*, 2006.

[20] E. M. Voorhees and D. Harman, editors. *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, Cambridge, MA, 2005.