News Recommendations using CF-IDF

Frederik Hogenboom Flavius Frasincar Uzay Kaymak Franciska de Jong

Erasmus University Rotterdam
P.O. Box 1738, NL-3000 DR, Rotterdam, the Netherlands
{*fhogenboom, frasincar, kaymak, fdejong*}*@ese.eur.nl*

The full version of this paper, entitled NEWS PERSONALIZATION USING THE CF-IDF SEMANTIC RECOMMENDER appeared in: Proceedings of the International Conference on Web Intelligence, Mining and Semantics 2011 (WIMS 2011), ACM, 2011

Abstract

Most of the traditional recommendation algorithms are based on TF-IDF, a term-based weighting method. This paper proposes a new method for recommending news items based on the weighting of the occurrences of references to concepts, which we call Concept Frequency-Inverse Document Frequency (CF-IDF). In an experimental setup we apply CF-IDF to a set of newswires in which we detect $1,167$ instances of a set of 65 concepts from a domain ontology. The proposed method yields significantly better results with respect to accuracy, recall, and F_1 than the TF-IDF method we use as a basis for comparison.

1 Introduction

In today's data intensive world, most people experience (or suffer from) an information overload. Recommender systems lend a hand in distinguishing between interesting and non-interesting products, movies, games, hotels, news articles, etcetera. Using, for example, user preferences or characteristics captured in user profiles based on user input or derived from browsing behavior, recommendations can be made.

A commonly used measure in recommender systems is TF-IDF [4], i.e., Term Frequency-Inverse Document Frequency. A major drawback of TF-IDF is that its performance decreases as documents get larger [2]. Lately, Semantic Web technologies have been developed that aid in finding key concepts in a text. We hypothesize that through the use of semantics, a lot of noise caused by non-meaningful terms would be reduced. Therefore, we propose Concept Frequency-Inverse Document Frequency (CF-IDF), which is analogous to TF-IDF, but instead of counting term frequencies, we count frequencies of specific concepts.

Although some related work has been done [1, 5], the performance of the proposed semantic methods for recommendations has not been thoroughly compared with TF-IDF. This paper presents CF-IDF, which is tested in a news recommendation system: Athena, i.e., an extension to the Hermes [3] news processing framework. Section 2 discusses the recommendation process in more detail, and Section 3 evaluates the proposed method. Finally, Section 4 presents our conclusions and future work directions.

2 CF-IDF

Currently, recommendations are often made based on TF-IDF. First, for each document in the corpus stop words are removed and the remaining words (terms) are stemmed to their roots. Then, term frequencies (i.e., importance of a term within a document) are multiplied with the inverse document frequencies (i.e., the inverse of the general importance of a term in a set of documents) to obtain a document term importance. Hence, the document term importance increases direct proportionally to the number of times a term appears in the document, but inverse proportionally with the frequency of the word in the corpus. In our proposed CF-IDF recommender, we use ontology concepts instead of terms in documents. These concepts are found using natural language processing pipelines [3].

When recommending news in Athena, we use a user profile that consists of a subset of the concepts and relations stored within an ontology. The user profile is constructed by keeping track of the articles a user reads and by extracting the most frequent terms and concepts. Each article is represented as a set containing all appearing terms (TF-IDF) or concepts (CF-IDF). Then, for each article, TF-IDF and CF-IDF weights are calculated. Weights of a new article are compared to the user profile using cosine similarity, resulting in a ranked list of possibly interesting news items according to a constructed user profile.

3 Evaluation

For evaluation purposes, we implemented TF-IDF and CF-IDF as a user profiling and recommendation plug-in in the Hermes News Portal (HNP), the implementation of the Hermes framework. Hermes provides a semantic-based approach for retrieving news items related, directly or indirectly, to the concepts of interest from a domain ontology. HNP takes RSS feeds of news items as input and detects concepts from a domain ontology in news through an advanced natural language processing engine. The ontology, which is developed manually by domain experts, contains a small subset of commonly used, well-known, financial entities such as companies, products, currencies, etc., and these concepts have associated lexical representations. The ontology consists of 65 classes, 18 object properties, 11 data properties, and $1,167$ individuals.

For recommendation evaluation, we let 19 users browse 100 news articles and indicate the interestingness when keeping in mind a predefined preference for Microsoft, its products, and its competitors. We use 60 news articles for training (computing the user profile) and 40 news articles for testing. Then, we let both recommenders determine the similarity with the user profile for each news item, using a cutoff value for interestingness. For the optimal threshold value of 0.4, our results show that CF-IDF outperforms TF-IDF on various aspects. The higher accuracy ($+4.2\%$) indicates that the CF-IDF recommender is significantly performing better in classifying both interesting and uninteresting items correctly. Also, on recall ($+24.0\%$), the number of interesting news items being classified as interesting, the CF-IDF recommender performs significantly better. This result also shows in the F_1 measure ($+19.1\%$). The performance for precision and specificity is also higher for CF-IDF, but this improvement is not significant.

4 Conclusions

In this paper we have presented an alternative to the TF-IDF recommendation approach, CF-IDF, which uses the knowledge available in an ontology. CF-IDF outperforms TF-IDF significantly in terms of accuracy, recall, and F_1. Hence, using key concepts and their semantics instead of analyzing all terms could be beneficial for recommender systems. As future work, we would like to experiment with different stemmers, as well as with different weighting schemes (for both CF-IDF and TF-IDF) that show good performance in the literature (e.g., Okapi).

References

[1] Mustapha Baziz, Mohand Boughanem, and Salam Traboulsi. A Concept-Based Approach for Indexing Documents in IR. In *Actes du XXIIIème Congrès Informatique des Organisations et Systèmes d'Information et de Décision (INFORSID 2005)*, pages 489–504. HERMES Science Publications, 2005.

[2] Toine Bogers and Antal van den Bosch. Comparing and Evaluating Information Retrieval Algorithms for News Recommendation. In *ACM Conference on Recommender Systems 2007 (RecSys 2007)*, pages 141–144. ACM, 2007.

[3] Flavius Frasincar, Jethro Borsje, and Leonard Levering. A Semantic Web-Based Approach for Building Personalized News Services. *International Journal of E-Business Research (IJEBR)*, 5(3):35–53, 2009.

[4] Gerard Salton and Chris Buckley. Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing and Management*, 24(5):513–523, 1988.

[5] Linyuan Yan and Chunping Li. A Novel Semantic-based Text Representation Method for Improving Text Clustering. In *3rd Indian International Conference on Artificial Intelligence (IICAI 2007)*, pages 1738–1750, 2007.