Bing-SF-IDF+: Semantics-Driven News Recommendation

Frederik Hogenboom
fhogenboom@ese.eur.nl

Michel Capelle
michelcapelle@gmail.com

Marnix Moerland
marnix.moerland@gmail.com

Flavius Frasincar
frasincar@ese.eur.nl

Erasmus University Rotterdam
PO Box 1738, NL-3000 DR
Rotterdam, the Netherlands

ABSTRACT

Content-based news recommendation is traditionally performed using the cosine similarity and TF-IDF weighting scheme for terms occurring in news messages and user profiles. Semantics-driven variants such as SF-IDF additionally take into account term meaning by exploiting synsets from semantic lexicons. However, they ignore the various semantic relationships between synsets, providing only for a limited understanding of news semantics. Moreover, semantics-based weighting techniques are not able to handle – often crucial – named entities, which are usually not present in semantic lexicons. Hence, we extend SF-IDF by also considering the synset semantic relationships, and by employing named entity similarities using Bing page counts. Our proposed method, Bing-SF-IDF+, outperforms TF-IDF and SF-IDF in terms of F_1 scores and kappa statistics.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Information filtering, Relevance feedback*; I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods—*Representation languages*

Keywords

Recommender systems, semantics-driven recommendation, Bing

1. INTRODUCTION

The Web has become an important information source to many users. Due to the resulting information overload, many recommendation methods have been developed that operate on user profiles. Traditionally, content-based recommenders use the Term Frequency – Inverse Document Frequency (TF-IDF) measure. When translating user profiles and news documents into TF-IDF weight vectors, and employing a similarity measure like cosine similarity, a document's relevance for a specific user is determined. A drawback is that text semantics are not considered, which could be overcome by using Web ontologies. However, these are often domain dependent and require continuous maintenance. Employing synonym sets (synsets) from general semantic lexicons (e.g., WordNet) eliminates the need for domain ontologies. Hence, in previous work [3], we introduced the Synset Frequency – Inverse Document Frequency (SF-IDF) recommender, using WordNet synsets instead of terms.

Up until now, we did not consider inter-synset relationships, though relationships as synonymy and hyponymy contribute to an improved level of interpretability. Also, named entities are ignored, although they could provide crucial information when constructing user profiles. The more a pair of entities co-occur on Web sites, the higher their similarity [1]. One could enhance existing semantics-based recommendation methods by employing similarities based on page counts gathered by Web search engines like Bing. Therefore, we extend SF-IDF by additionally considering WordNet synset semantic relationships and Bing page count-based named entity similarities. We evaluate our method, Bing-SF-IDF+, against the TF-IDF and SF-IDF baselines.

2. RELATED WORK

One of the most common recommendation approaches is TF-IDF, used with cosine similarities. The TF-IDF method is composed of the term frequency $tf(t, d)$ and inverse document frequency $idf(t, d)$, and operates on terms T in documents D. The term frequency measures the number of occurrences n of term $t \in T$ in document $d \in D$ expressed as a fraction of the total number of occurrences of all k terms in document d. The inverse document frequency expresses the occurrence of a term t in a set of documents D and is obtained by dividing the cardinality of D by the number of documents d containing term t, and then taking the logarithm of that quotient. We obtain $tf\text{-}idf(t, d)$ by multiplying $tf(t, d)$ and $idf(t, d)$. Next, for every term t in document d, the TF-IDF value is computed and stored in a vector $A(d)$. The user profile is defined as the vector corresponding to the document obtained by concatenating all the previously visited articles.

A TF-IDF variant is the Synset Frequency – Inverse Document Frequency (SF-IDF) [3], which makes use of synonym sets (synsets) from a semantic lexicon instead of terms. These synsets are obtained after performing word sense dis-

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2577310.

ambiguation using an adapted Lesk algorithm. After every unread document has been assigned a value representing its cosine similarity with the user profile using vectors of term weights (TF-IDF) or synset weights (SF-IDF), the unread news items with a similarity value higher than a cut-off value are recommended to the user.

3. BING-SF-IDF+ RECOMMENDATION

Like most semantics-based news recommendation methods, Bing-SF-IDF+ operates on a user profile, consisting of read news items, which is updated upon reading previously unseen news items. For every unread news item, a similarity score between the news article and the user profile is computed, which is a weighted average of two scores. The Bing component expresses similarities between named entities, and SF-IDF+ measures the synset similarities.

The Bing similarity score takes into account the named entities not occurring in a semantic lexicon, which are derived through a named entity recognizer. We describe an unread news item d_u and the user profile d_r using sets of named entities U and R. We construct a vector V containing all possible pairs of named entities from d_u and d_r. Next, we use search engine page counts of the named entity pairs for measuring pair similarities. For every pair (u, r) in V, we compute the Point-Wise Mutual Information (PMI) co-occurrence similarity [2]. Last, the Bing similarity score $sim_{Bing}(d_u, d_r)$ is defined as the average of the PMI similarity scores over all named entity pairs.

The SF-IDF+ similarity score takes into account sets of synonyms (synsets) of words and is based on SF-IDF. Similarities $sim_{sf\text{-}idf^+}(d_u, d_r)$ are the same cosine similarities as before, yet the score vectors $A(d)$ for a document or profile d are different. Now, not only directly occurring synsets are retrieved, but also the synsets from these concepts that are referred to by their semantical relationships. Similar to TF-IDF and SF-IDF, similarity scores are calculated, but now an additional weighting is applied depending on the relationships between synsets and their semantically related synsets. Weights are optimized using a genetic algorithm.

Last, Bing and SF-IDF+ similarity scores are normalized using min-max normalization between 0 and 1. The Bing-SF-IDF+ similarity score $sim_{Bing\text{-}sf\text{-}idf^+}(d_u, d_r)$ is computed by taking a weighted average of the normalized similarity scores using an optimized weight α.

4. RESULTS AND CONCLUSIONS

We evaluate Bing-SF-IDF+ against SF-IDF and TF-IDF using 100 news articles from a Reuters news feed on technology companies, annotated by 3 experts for their relevance with respect to 8 topics, using a minimum inter-annotator agreement of 66%. Methods are compared based on their F_1 score (harmonic mean of precision and recall) and kappa statistics (measuring whether the proposed classification is better than a random guess). Performances are evaluated on a test set (60%) for cut-off values ranging from 0 to 1 with an increment of 0.01. Significance is assessed using one and two-tailed two-sample paired Student t-tests with a significance level of 95%. We optimize the Bing-SF-IDF+ weights and α-value on a training set (40%) using a genetic algorithm, maximizing F_1 scores. The algorithm is executed with an optimized configuration using a population of 333, a mutation probability of 0.1, elitism of 50, and a maximum number of 1,250 generations. Experiments are run on the Lisa system, a large multi-core SARA Computing and Networking Services cluster computer.

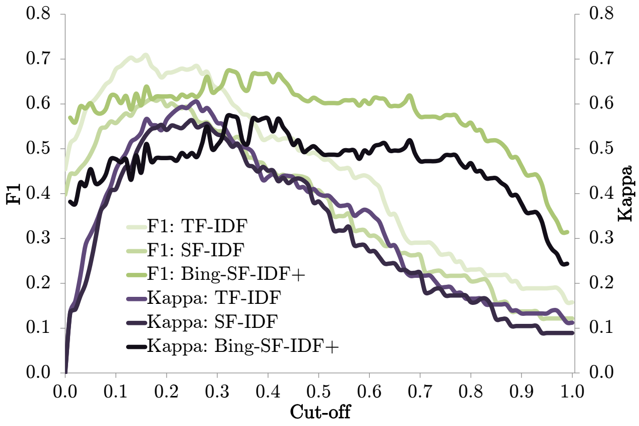

Figure 1: Experimental results.

As depicted in Figure 1, Bing-SF-IDF+ significantly outperforms SF-IDF and TF-IDF in terms of average F_1 scores over all topics and cut-off values, scoring 0.58 against 0.37 and 0.43, respectively. Also when comparing kappa statistics, SF-IDF and TF-IDF are outperformed by Bing-SF-IDF+, as the former two methods have averages of 0.32 and 0.34, respectively, and the latter method has an average of 0.47. Bing-SF-IDF+ score weight α is optimized to 0.48 (with a standard deviation of 0.27), giving a substantial weight to both Bing similarities and extended synsets incorporating semantic relationships. WordNet relationships that typically obtain high weights are ‘attribute’, ‘derivationally related form’, ‘derived from adjective’, ‘instance hyponym’, ‘substance holonym’, ‘member meronym’, and ‘member of this domain - usage’, which can be explained by the fact that these are semantically rich relations.

Acknowledgments

The authors are partially supported by the NWO Physical Sciences Free Competition project 612.001.009: Financial Events Recognition in News for Algorithmic Trading (FERNAT) and the Dutch national program COMMIT. We thank SARA Computing and Networking Services for their support in using the Lisa Computer Cluster.

5. REFERENCES

[1] D. Bollegala, Y. Matsuo, and M. Ishizuka. Measuring Semantic Similarity between Words Using Web Search Engines. In *16th Int. Conference on World Wide Web (WWW 2007)*, pages 757–766. ACM, 2007.

[2] G. Bouma. Normalized (Pointwise) Mutual Information in Collocation Extraction. In C. Chiarcos, R. E. de Castilho, and M. Stede, editors, *Biennial GSCL Conference 2009 (GSCL 2009)*, pages 31–40. Gunter Narr Verlag Tübingen, 2009.

[3] M. Capelle, M. Moerland, F. Frasincar, and F. Hogenboom. Semantics-Based News Recommendation. In *2nd International Conference on Web Intelligence, Mining and Semantics (WIMS 2012)*. ACM, 2012.