Erasmus University Rotterdam

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

Introduction

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 to **Bing-SF-IDF+** by also considering the synset semantic relationships, and by employing named entity similarities using Bing page counts.

Classic Recommendation

Classic news recommenders commonly use a Term Frequency – Inverse Document Frequency (**TF-IDF**) weighting scheme. Here, **scores** increase proportionally to the number of times a specific **term** (word) appears in a document, but are scaled with the frequency of a word in the full set of known documents (the corpus). Hence, terms that occur relatively often in a document, but not in the rest of the corpus, are assigned high scores, highlighting their importance, while common terms receive low scores.

Usually, term scores are computed for the **user profile** (i.e., a collection of read documents) and for a **new document**. These scores are subsequently compared using a similarity measure, such as the **cosine similarity**. Documents displaying a high term score similarity with the user profile (i.e., above a certain specified **cut-off** value) are recommended.

A recent variant is the Synset Frequency – Inverse Document Frequency (**SF-IDF**), which takes into account semantics by operating on synonym sets (**synsets**) from a **semantic lexicon**, instead of terms. These synsets are obtained after performing word sense **disambiguation** (using **SSI**, **Lesk**, etcetera).

Bing-SF-IDF+ Recommendation

In essence, **Bing-SF-IDF+** recommendation is similar to the synset-based SF-IDF method, as synset score similarities between the user profile and a new document are computed. Additionally, however, named entities (derived through a named entity recognizer) and synset relationships are considered. The method computes a weighted average of the entity-based and relationship-based similarity scores.

Named entities that have been identified in the user profile and a new document, and that do not occur in a semantic lexicon, are submitted in pairs to the **Bing** search engine. Page counts are evaluated in order to compute the Point-Wise Mutual Information (**PMI**) entity co-occurrence similarity, measuring the discrepancy between the probability of the entities' coincidence, given their joint distribution and their individual distributions. Frequently co-occurring entities are assumed to be highly similar, and hence contribute to the similarity between the user profile and a new document.

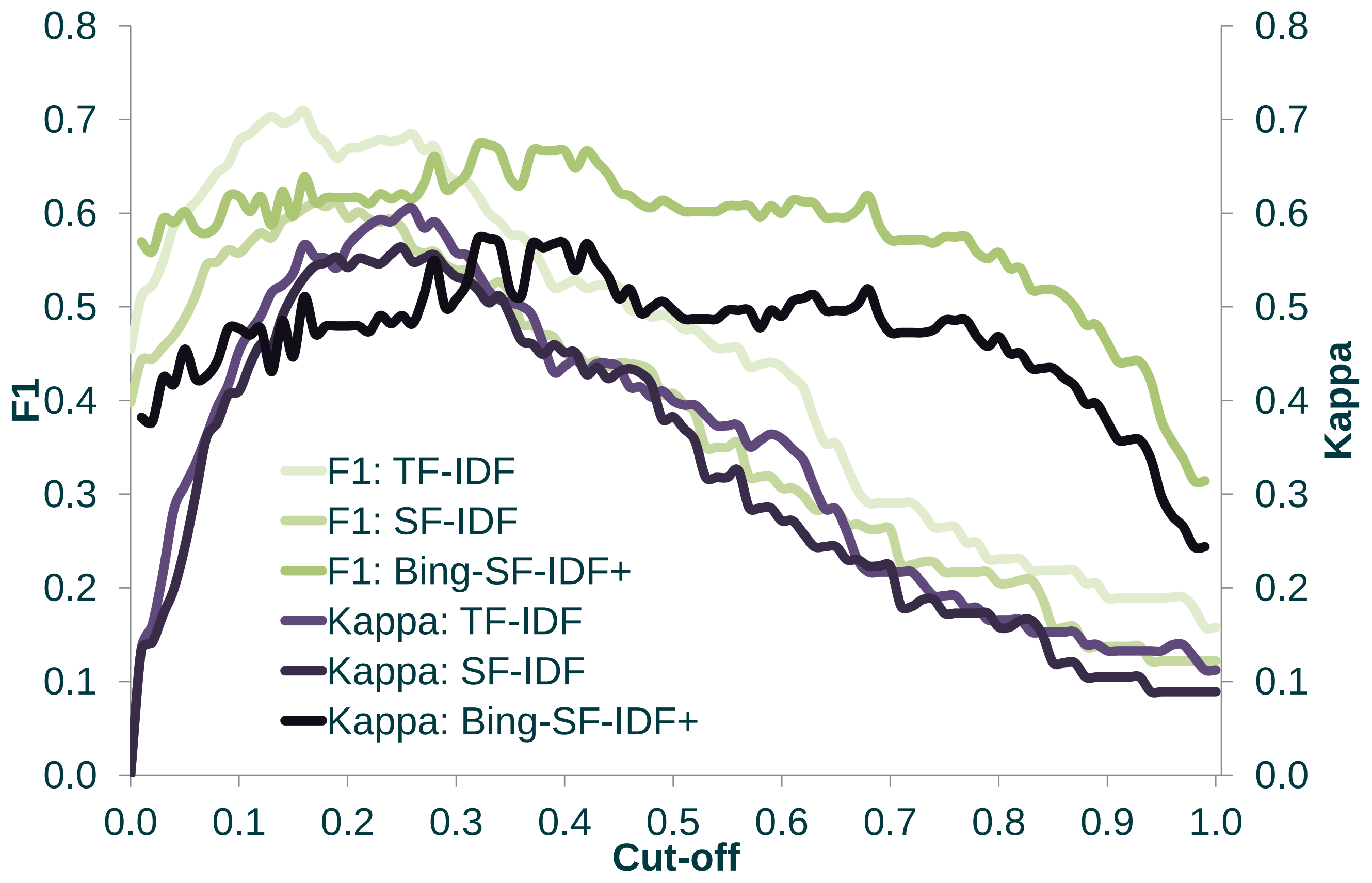

Fig. 1. Experimental results.

For identified **synsets**, the similarities are computed identically to SF-IDF, yet the score vectors contain not only the directly occurring synsets, but also the synsets from these concepts that are referred to by their **semantical relationships**. Additional weighting is applied depending on the relationships, and weights are optimized using a genetic algorithm.

Results & Conclusions

Experiments on 100 **news documents** from a **Reuters** news feed on technology companies, annotated by 3 experts for their relevance with respect to 8 topics, show that Bing-SF-IDF+ significantly **outperforms** SF-IDF and TF-IDF in terms of average **F1** scores over all topics and cut-off values. As depicted in Fig. 1, Bing-SF-IDF+ obtains a score of 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. Results are optimal when giving an equal weight to both Bing similarities and extended synsets incorporating semantic relationships.

Acknowledgement

This publication was partially supported by the Dutch Organization for Scientific Research (NWO) Physical Sciences Free Competition Project 612.001.009: Financial Events Recognition in News for Algorithmic Trading (FERNAT) and the Dutch national program COMMIT.

Contact

Frederik Hogenboom
Postal: Erasmus University Rotterdam
P.O. Box 1738
NL-3000 DR Rotterdam
The Netherlands
Phone: +31 (0)10 408 1262
E-Mail: fhogenboom@ese.eur.nl
Web: http://people.few.eur.nl/fhogenboom/

Frederik Hogenboom, Michel Capelle, Marnix Moerland, Flavius Frasincar