Knowledge-Based Systems 54 (2013) 216–229

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Semantic smoothing for text clustering

Jamal A. Nasir [a], Iraklis Varlamis [b,*], Asim Karim [a], George Tsatsaronis [c]

[a] KADE Lab, Lahore University of Management Sciences, Pakistan
[b] Department of Informatics and Telematics, Harokopio University of Athens, Greece
[c] BIOTEC, Technical University of Dresden, Germany

ARTICLE INFO

Article history:
Received 26 February 2013
Received in revised form 5 September 2013
Accepted 7 September 2013
Available online 24 September 2013

Keywords:
Text clustering
Semantic smoothing kernels
WordNet
Wikipedia
Generalized vector space model kernel

ABSTRACT

In this paper we present a new *semantic smoothing* vector space kernel (*S-VSM*) for text documents clustering. In the suggested approach semantic relatedness between words is used to smooth the similarity and the representation of text documents. The basic hypothesis examined is that considering semantic relatedness between two text documents may improve the performance of the text document clustering task. For our experimental evaluation we analyze the performance of several semantic relatedness measures when embedded in the proposed (*S-VSM*) and present results with respect to different experimental conditions, such as: (i) the datasets used, (ii) the underlying knowledge sources of the utilized measures, and (iii) the clustering algorithms employed. To the best of our knowledge, the current study is the first to systematically compare, analyze and evaluate the impact of semantic smoothing in text clustering based on '*wisdom of linguists*', e.g., *WordNets*, '*wisdom of crowds*', e.g., *Wikipedia*, and '*wisdom of corpora*', e.g., large text corpora represented with the traditional *Bag of Words* (*BoW*) model. Three semantic relatedness measures for text are considered; two knowledge-based (*Omiotis* [1] that uses *WordNet*, and *WLM* [2] that uses *Wikipedia*), and one corpus-based (*PMI* [3] trained on a semantically tagged *SemCor* version). For the comparison of different experimental conditions we use the *BCubed F-Measure* evaluation metric which satisfies all formal constraints of good quality cluster. The experimental results show that the clustering performance based on the *S-VSM* is better compared to the traditional *VSM* model and compares favorably against the standard *GVSM* kernel which uses word co-occurrences to compute the latent similarities between document terms.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

Document clustering plays an important role in indexing, retrieval, browsing and mining of large and high dimensional text data. Document clustering algorithms aim at organizing documents into meaningful groups that contain highly similar documents, and which are distant from the documents of other groups [4]. For this purpose, they rely on document similarity or distance measures, with which they typically compare pairs of text documents. Therefore, similarity measures play a crucial role in the task of document clustering. The performance of similarity measures in data mining tasks depends on the type of data, on the particular domain, on the dataset and on the nature of the examined task. In the case of document clustering, the textual data have usually large volume, they are high-dimensional, and carry also semantic information, i.e., meaning conveyed by the text terms. Therefore, the clustering algorithm and the similarity measures that are employed for the task should be able to address these parameters effectively.

In the task of document clustering documents are typically represented by their terms. Terms are either single words or composite (multi-word terms), which form as a whole the language vocabulary of the underlying text corpus. Terms of either category are usually associated with a positive real value acting as a weight for the respective term. Furthermore, the weight of each term corresponds to its importance/relevance to the document it appears in.

More formally, given a collection of documents D, the vocabulary V of D may be defined as the set of all distinct terms appearing in D. For each term t_i of a document $d_j \in D$, the weight $w_{ij} > 0$ of the t_i in d_j may be computed, usually, through a measure that takes into account the frequency of occupancies of t_i in d_j. This representation is widely known as the *Vector Space Model* (*VSM*) [5].

VSM is very simple and commonly used; yet, it has several limitations. Its main restriction is that it assumes independency between the vocabulary terms and ignores all the conceptual relations between terms that potentially exist. As a consequence, two terms that are semantically close, e.g., synonyms, are treated differently. Furthermore, polysemous terms, i.e., terms with multiple meanings, are considered the same in all contexts they

* Corresponding author. Tel.: +30 2109549295; fax: +30 2109549405.
E-mail addresses: varlamis@hua.gr, varlamis@gmail.com (I. Varlamis).

0950-7051/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.knosys.2013.09.012

appear. For example the term '*bank*' may have the meaning of a *financial institution* when it appears in a context related to economy, or the meaning of a *river side* when it appears in a context that refers to landscapes or geographical locations. In principle a document contains usually more terms that are '*general terms*', i.e., that may appear in all clusters, than '*cluster dependent terms*', i.e., '*core terms*' [6] that characterize the documents of a single cluster. *VSM* cannot consider that differentiation as it cannot examine similarities between terms that have different surface strings. For the *VSM* model, the similarities between documents and the similarities between a document and the cluster centroid are only based on the matched term strings. Hence, the need of smooth semantically the *VSM* model, i.e., by employing semantic smoothing *VSM* kernels, arises. This embedding may increase the importance of *core words* by considering the terms' relations, and in parallel downsize the contribution of *general terms*, leading to better text clustering results.

In this article, we propose a novel semantic smoothing *VSM* kernel (*S-VSM*), which smooths the *VSM* representation with the semantic similarity between terms.[1] The proposed *S-VSM* allows any semantic similarity or relatedness measure to be employed, both measures that use linguistic resources, e.g., knowledge bases, ontologies, and thesauri, but also measures that are based on statistical information extracted from the analysis of large text corpora. Hence, the first advantage of the suggested solution is that it offers a very flexible kernel that can be applied within any domain or with any language. To showcase the wide applicability of the suggested kernel, for the purposes of this work we examine the embedding of three novel semantic relatedness measures into the *S-VSM*; the first employs the *WordNet*-based similarity measure of *Omiotis* [1], the second is *Wikipedia*-based and employs the measure of Milne and Witte [2], and the third is based on statistical analysis of text corpora and uses a *Pointwise Mutual Information* similarity measure for the computation of terms' similarity [3].

The second advantage of the suggested solution is the ability of the *S-VSM* to perform much better than the *VSM* in the task of text clustering. In addition, an extension of the *S-VSM* that we propose, namely the *top-k S-VSM*, which considers only the top-*k* semantically related terms, does not only perform better than the *VSM*, but it also conducts the task of text clustering very efficiently in terms of time and space complexity. The proposed *S-VSM* and its extension are evaluated on five datasets: (1) *Reuters-Transcribed-set*,[2] (2) *R8 Reuters-21578*,[3] (3) *4 Universities Data Set* (*WebKB*),[4] (4) *Email-1431*,[5] and (5) *Movie Reviews*.[6] To evaluate *S-VSM* and *top-k S-VSM* we use both *agglomerative* and *partitional* clustering for conducting the experiments, and two baselines; the traditional Bag of Word (*BoW*) model which uses the *VSM* model for document representation, and the standard *Generalized Vector Space Model* kernel (*GVSM*), which considers the term-to-document matrix to compute latent similarities between terms based on their co-occurrence.

The clustering results show significant improvements in the clustering accuracy when *S-VSM* and *top-k S-VSM* are used, compared to the performance of the two aforementioned baselines. In addition, we provide a thorough analysis on the effect of the number of the top-*k* semantically related terms used for the smoothing, which, to the best of our knowledge, is conducted for the first time in the bibliography, and gives important insights on how the semantic smoothing can be optimized computationally.

This work capitalizes on our previous work on semantic kernels [7]. The main contributions of the current work, which differentiate it from our former work on *S-VSM* kernels and expand it, can be summarized in the following:

1. Extension of the $S - VSM$ to embed only the top-*k* semantically related terms.
2. Application to the task of text clustering.
3. An extended and thorough evaluation in text clustering, using a large variety of text datasets, employed clustering algorithms, and evaluation metrics.
4. Comparative evaluation against the standard *GVSM* kernel and the semantic kernel presented in [8], which shows that the suggested expanded *S-VSM* performs favorably against these two approaches.

The rest of the paper is organized as follows: Section 2 discusses the related work in the field of semantic smoothing kernels, with emphasis to the task of text clustering. Section 3 provides preliminary related information. Section 4 introduces the semantic smoothing kernel (*S-VSM*) and its *top-k* extension. Section 5 presents the experimental setup, and Section 6 presents and analyzes the experimental results. Finally, we conclude the paper in Section 7 and provide a discussion on the possible expansions of the current work.

2. Related work

The idea of using background knowledge or gathered statistical information from large text corpora analysis in order to compute text similarity is well studied in the past [9,10], and there exist many research works that introduce efficient similarity or relatedness measures between terms. With regards to works that employ such measures for document clustering, *WordNet* is one of the most widely used lexical thesauri [11,12]. In principle, research works in document clustering, but also in text retrieval, that incorporate semantics in the *VSM* representation can be classified in three categories, depending on the type of information or the features used to index the document terms and expand the index with additional features: (i) embedding of concept features, (ii) embedding of multi-word phrases, and (iii) employing semantic kernels to embed semantically related terms or semantic relation information between terms to the documents' representation; the semantic similarity and relations may be retrieved from a word thesauri or ontology, or may be computed based on statistical analysis of a large text corpus.

Works in the first category, e.g., [11], use conceptual features to improve the clustering performance. *WordNet* is typically used as a background knowledge to obtain concept features, which are defined as set of words that describe the same high level concept; for example *penny*, *nickel*, *dime* and *quarter* describe the concept *coin*. The weights of both concepts and terms are employed to represent documents, usually leading to a type of *hybrid* document representation in the vector space, that contains both concepts, i.e., meanings, but also concrete terms. Such representations were also applied in the past in text retrieval, with mixed performance outcome[7] [14]. Another recent representative example of a work

[1] Though there are slight conceptual differences between the terms '*semantic similarity*' and '*semantic relatedness*', for the purposes of this work this differentiation is not important. Therefore, the two terms might be used interchangeably for the remaining of the paper.

[2] Available for download from http://archive.ics.uci.edu/ml/datasets/Reuters+Transcribed+Subset.

[3] Available for download from http://web.ist.utl.pt/acardoso/datasets/datasets.zip.

[4] From the *WebKB* project, available for download from http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/webkb-data.gtar.gz.

[5] Available for download from http://cogsys.imm.dtu.dk/toolbox/nmf/email.zip.

[6] Available for download from http://www.cs.cornell.edu/people/pabo/movie-review-data/review_polarity.tar.gz.

[7] Usually such hybrid representations improve the recall in information retrieval systems, due to the expansion in the concepts' dimension, but might drop precision due to the difficulty of transiting from terms to concepts, with the task of word sense disambiguation having major role and innate limitations [13].

Download English Version:

https://daneshyari.com/en/article/6862733

Download Persian Version:

https://daneshyari.com/article/6862733

Daneshyari.com