Expert Systems With Applications 53 (2016) 138–148

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Microblog semantic context retrieval system based on linked open data and graph-based theory

Fahd Kalloubi*, El Habib Nfaoui, Omar El beqqali

LIIAN laboratory, Sidi Mohamed Ben Abdellah University, Fez, Morocco

ARTICLE INFO

Keywords:
Information retrieval
Semantic similarity
Linked open data
DBpedia
Named entity linking
Graph centrality

ABSTRACT

Microblogging platforms have emerged as large collections of short documents. In fact, the provision of an effective way to retrieve short text presents a significant research challenge owing to several factors: creative language usage, high contextualization, the informal nature of micro blog posts and the limited length of this form of communication. Thus, micro blogging retrieval systems suffer from the problems of data sparseness and the semantic gap. This makes it inadequate to accurately meet users' information needs because users compose tweets using few terms and without query terms inside; thus, many relevant tweets will not be retrieved. To overcome the problems of data sparseness and the semantic gap, recent studies on content-based microblog searching have focused on adding semantics to micro posts by linking short text to knowledge bases resources. Moreover, previous studies use bag-of-concepts representation by linking named entities to their corresponding knowledge base concepts. However, bag-of-concepts representation considers only concepts that match named entities and supposes that all concepts are equivalent and independent. Thus, in this paper, we present a graph-of-concepts method that considers the relationships among concepts that match named entities in short text and their related concepts and contextualizes each concept in the graph by leveraging the linked nature of DBpedia as a Linked Open Data knowledge base and graph-based centrality theory. Furthermore, we propose a similarity measure that computes the similarity between two graphs (query-tweet) by considering the relationships between the contextualized concepts. Finally, we introduce some experiment results, using a real Twitter dataset, to expose the effectiveness of our system.

1. Introduction

Microblogging platforms allow users to exchange short texts, such as tweets and user statuses in friendship networks. Microblogging has emerged as one of the primary social media platforms for users to post short messages and content of interest. Twitter is one of the most popular microblog service providers. In fact, it has attracted more than 500 million registered users and publishes 340 million tweets per day,[1] and many queries are issued each day (more than 1.6 billion search queries[2] in Twitter); however, determining the subject of an individual micro post can be nontrivial owing to several factors: creative language usage, the highly contextualized, informal nature of microblog posts, and the limited length of this form of communication. Therefore, these factors make micro blogging streams an invaluable sources for many types of analyses, including online reputation management, news and trend detection; targeted marketing and customer services (Boyd, Golder, & Lotan, 2010; Kwak, Changhyun, Hosung, & Sue, 2010; Manos, de Rijke, & Weerkamp, 2011; Xiangmin & Lei, 2013); these applications mainly analyze and utilize the wisdom of the crowds as a source of information rather than relying on individual tweets. Searching and mining microblog streams offer interesting technical challenges in many microblog search scenarios, and the goal is to determine what people are saying about concepts such as products, brands, and persons (Brendan, Ramnath, Bryan, & Noah, 2010). Microblogging retrieval systems suffer from the problems of data sparseness and the semantic gap owing to the length of microblog posts and their high contextualization. This makes it inadequate to accurately meet users' information needs because users compose tweets using different terms and without query terms inside; thus, many relevant tweets will not be retrieved. Most current microblogging Information Retrieval (IR) systems rely on a term-based model such as TF-IDF, BM25 and the probabilistic model (Lau, Li, & Tjondondronegoro, 2011).

* Corresponding author. Tel.: +212661860460.
E-mail addresses: fahd.kalloubi@usmba.ac.ma (F. Kalloubi), elhabib.nfaoui@usmba.ac.ma (E.H. Nfaoui), omar.elbeqqali@usmba.ac.ma (O. El beqqali).

[1] http://techcrunch.com/2012/07/30/analyst-twitter-passed-500m-users-in-june-2012-140m-of-them-in-us-jakarta-biggest-tweeting-city/ (last access: December 2015).

[2] http://techcrunch.com/2011/06/01/new-twitter-search-relevant/ (last access: December 2015).

http://dx.doi.org/10.1016/j.eswa.2016.01.020
0957-4174/© 2016 Elsevier Ltd. All rights reserved.

Term-based models are efficient in terms of computation performance, and the maturity of term weighting theories has made these models rampant. However, term-based approaches often suffer from the problems of polysemy and synonymy and are very sensitive to term use variation. Topical features such as phrases and named entities (e.g., person, location and proper nouns) are also often neglected (Lau et al., 2011). This problem is even more evident in microblogs owing to the amount of noise, which causes poor retrieval performance. To overcome this problem, many research efforts have been conducted to understand and model the semantics of individual microblog posts. Linking free text to knowledge resources, however, has received an increasing amount of attention in recent years. Starting from the domain of named entity recognition (NER), current approaches establish links not just to entity types but to the actual entities themselves (Meij, Weerkamp, & de Rijke, 2012; Xiaohua, Shaodian, Furu, & Ming, 2011). With more than 3.5 million articles, Wikipedia has become a rich source of knowledge and a common target for linking; automatic linking approaches using Wikipedia have achieved considerable success (He, de Rijke, Sevenster, van Ommering, & Qian, 2011; Meij, Bron, Hollink, Huurnink, & de Rijke, 2011). DBpedia as a central Linked Open Data dataset is a knowledge base created from Wikipedia by converting structured information (e.g., infobox) to a Resource Description Framework (RDF) data model.

The main contribution of the Semantic Web as a new form of Web content is the provision of meaning to computers with the utilization of ontology as a source of knowledge representation (Berners-Lee, Hendler, & Lassila, 2001). The Resource Description Framework (RDF) is introduced as an underlying framework to use ontology in the Web environment. The RDF data model treats each piece of information as a triple: subject–property–object (Lassila & Swick, 1999). The application of RDF for data representation has become a very popular means of representing data on the Web. Over time, more attention has been paid to it, and the term Linked Open Data (LOD) has been used to describe the network of data sources based on RDF triples for information representation (Shadbolt, Hall, & Berners-Lee, 2006). The main contribution of LOD contrary to hypertext Web is that entities from different sources/locations are linked to other related entities on the Web by the use of a Unified Resource Identifier (URI). This enables one to view the Web as a single global data space (Bizer, Heath, & Berners-Lee, 2009). In other words, hypertext Web connects documents in a naive way. However, in the Web of LOD, single information items are connected. As a result, DBpedia allows for better representation of structured data in a machine-understandable way.

We propose a graph-of-concepts method that considers the relationships between concepts and their related concepts and contextualizes each concept in the graph by leveraging the linked nature of DBpedia as a knowledge base. Furthermore, we propose a similarity measure that computes the similarity between two graphs (query-tweet). Our similarity measure considers the overlapping between named entities, which have been shown to obtain the best results in microblog searching (Tao, Abel, Hauff, & Houben, 2012a) and the relationships between the contextualized concepts in the graph.

The remainder of this paper is organized as follows. In Section 3, we present a method for enriching and adding semantic to micro posts by processing tweets and linking the entities extracted to LOD concepts using DBpedia as a knowledge base. This will help us represent micro posts as graphs-of-concepts by leveraging the linked nature of DBpedia to define our semantic context similarity measures. Equally important, we present our approach to contextualize all entities extracted by using graph-based centrality scoring. In Section 4, we introduce our algorithms for semantic context retrieval over tweets using the constructed graph-based context. Finally, in Section 5, we evaluate our system using a real dataset harvested from Twitter.

2. Related work

In this section, we review related works for adding semantic information to tweets and Information Retrieval systems over tweets.

2.1. Enriching and adding semantics to tweets

Linking text to a knowledge structure has received a great deal of attention, especially in the social Web because of the lack of semantics in such a complex structure. A rampant approach of linking text to concepts is to perform lexical matching between parts of text and the concept titles (Mendes, Passant, Kapanipathi, & P. Sheth, 2010). However, lexical matching suffers from many drawbacks, including ambiguity (polysemy and synonymy) and possible lack of specificity (less “meaningful” concepts are identified). Short texts have the characteristics of sparsity and noisiness owing to their limited length. Thus, when using the “bag of words” model to represent short text, contextual information is neglected and hence often leads to synonymy and polysemy problems (Tang, Wang, Gao, Hu, & Liu, 2012). To overcome the problem of data sparseness and the semantic gap in short text, various approaches have been proposed for adding semantics to text contained in tweets. (Somnath, Krishnan, & Ajay, 2007) developed a method to enrich short text representation with additional features from Wikipedia. This method used only the titles of Wikipedia articles as additional external features; it showed improvement in the accuracy of short text clustering. (Xiaohua et al., 2011) focused on NER in tweets and used a semi-supervised learning framework to identify four types of entities.

Meij et al. (2012)) proposed an approach to link n-grams to Wikipedia concepts based on various features. Their approach is divided into two steps; in the former, they generate a ranked list of candidate concepts for each n-gram in a tweet by applying a various type of features (*n*-grams features, concept features, tweet features). In the latter, they aim to improve precision by applying supervised machine learning. However, in our unsupervised approach, we consider only concept features by using different properties and contextualizing each concept by leveraging the linked nature of DBpedia as knowledge base to construct a weighted graph-of-concepts representation that depicts the context of each tweet by performing semantic linking and named entity resolution together, unlike their approach in which they suppose that all concepts are equal.

Abel, Gao, Houben, and Tao (2011) presented an approach that aims to contextualize tweets. After adding context, the authors use the tweets to profile Twitter users. Their approach is based on semantic enrichment with the news article’s content. Finally, the semantically enriched tweets are used for user modeling. Pertaining to the semantic enrichment, the authors use OpenCalais. Our approach proposed in Section 3 differs from their approach in the sense that we assume that the tweets are not related to news article, which makes our approach more general.

Tang et al. (2012) presented a framework for enriching short text for clustering purposes in which they perform multi-language knowledge integration and feature reduction simultaneously through matrix factorization techniques.

Mendes et al. (2010) proposed Linked Open Social Signals, a framework that includes annotating tweets with information from Linked Data. Their approach is rather straightforward and involves either looking up hashtag definitions or lexically matching strings to recognize (DBpedia) entities in tweets.

Kwak, Changhyun, Hosung, and Sue (2010) showed that hashtags are good indicators to detect events and trending topics, and