Expert Systems with Applications 38 (2011) 1575–1582

Contents lists available at ScienceDirect

Expert Systems with Applications

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

A language model approach for tag recommendation

Ke Sun [a,*], Xiaolong Wang [a,**], Chengjie Sun [a,1], Lei Lin [a,b,1]

[a] School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China
[b] Department of Control Science and Engineering, Harbin Institute of Technology, Harbin 150001, China

ARTICLE INFO

Keywords:
Tag recommendation
Language model for tag recommendation

ABSTRACT

Tags are user-generated keywords for entities. Recently tags have been used as a popular way to allow users to contribute metadata to large corpora on the web. However, tagging style websites lack the function of guaranteeing the quality of tags for other usages, like collaboration/community, clustering, and search, etc. Thus, as a remedy function, automatic tag recommendation which recommends a set of candidate tags for user to choice while tagging a certain document has recently drawn many attentions. In this paper, we introduce the statistical language model theory into tag recommendation problem named as language model for tag recommendation (LMTR), by converting the tag recommendation problem into a ranking problem and then modeling the correlation between tag and document with the language model framework. Furthermore, we leverage two different methods based on both keywords extraction and keywords expansion to collect candidate tag before ranking with LMTR to improve the performance of LMTR. Experiments on large-scale tagging datasets of both scientific and web documents indicate that our proposals are capable of making tag recommendation efficiently and effectively.

1. Introduction

Tags are user-generated keywords for entities to organize them with their common attribute. In opposite to the predefined organization style "taxonomy", this tagging based self generated organization style is called "folksonomy", which differs from the taxonomy of forcing the entities into the predefined categories, but presents a more flexible style by allowing people to freely annotate entities with their own keywords. Recently tags have been used as a popular way to allow users to contribute metadata to large corpora on the web by many famous websites (e.g. Delicious, Flickr). Its advantages make it suitable for organizing the web objects which changes rapidly in their distribution or types.

Although tagging is easy to perform and has many advantages, there are also some drawbacks. Golder and Huberman (2006) identified three major problems with current tagging systems:

- Polysemy. In tag system, polysemy refers to instances where a single tag can have multiple meanings. For example, the famous Company: "Apple" versus Fruit: "Apple".
- Synonymy, which means multiple tags having the same meaning. For example, the "news" versus "current events" or the misspelling problem likes "Nokia" versus "Nokea".
- Level variation, refers to the phenomenon of users tagging content at different levels of abstraction. Content can be tagged at a "basic level" or at varying levels of specificity which is often based on the tag poster's expertise or requirements. For example, given an entity like Google, normal users may use "search engine", "famous web site" or "good se", etc., to mark it, but researchers concerning on the IR techniques made using keywords from more academically fields like "page rank strategy", "map reduce" or "distributed indexing system".

These problems are caused by the lacking of clear functional pressure to make tagging consistent, stable and complete. And consequently, the collected tags are hard to be used in applications dealing with collaboration/community, clustering, and search.

In order to tackle these problems, the tag recommendation system has been recently proposed. It reminds the users of the alternative tags with less polysemy, synonymy problems from different abstraction levels. User could be reminded and some more suitable tags maybe selected easily. For example, when user wants to tag a document like "what is the single chip? what does it for?", the recommendation system will generate a list of recommendation tags based on the given document, such as "computer", "single chip", "hardware" and "electronic engineer". Also, because the tags can mostly be "selected" instead of "typing", the misspelling problem could be controlled. The recommendation system could not only help tackle the level variation problem by encouraging user fulfills more useful tags from different abstraction levels, but also improve the quality of posting tags by proposing candidate tags with less polysemy and synonymy problems.

* Corresponding author. Tel./fax: +86 451 86413322, mobile: +86 13946162826.
** Corresponding author. Tel.: +86 451 86413322.
E-mail addresses: ksun@insun.hit.edu.cn (K. Sun), wangxl@insun.hit.edu.cn (X. Wang), cjsun@insun.hit.edu.cn (C. Sun), linl@insun.hit.edu.cn (L. Lin).
[1] Tel.: +86 451 86413322.

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.07.075

In this paper, we focus on the tag recommendation problem for documents. By converting the tag recommendation problem into the problem of retrieving a set of tags relevant to the given document, a language model approach for tag recommendation (LMTR) is proposed. The statistical language model has been used for many nature language processing applications such as speech recognition, part-of-speech tagging, and syntactic parsing. In 1998, Ponte and Croft (Ponte & Croft, 1998) first introduced the language model approach for information retrieval by ranking the retrieved documents based on the probabilities of generating a query from the corresponding language models of these documents. Although the language model theory has been studied for years in many domains, but to the best of our knowledge, this is the first effort of introducing the statistical language model theory into the tag recommendation problem.

Our contribution focuses on the tag recommendation algorithm in documents. Specially, we (a) propose a novel tag recommendation framework based on the statistical language model theory, and (b) two expansion methods based on keywords extraction and keyword expansion theories for improving the tagging speed and performance are proposed also. Effectiveness and efficiency are both carefully analyzed for these proposals.

The reminder paper is organized as follows: In Section 2, we first survey the related work about tag recommendation and language modeling. In Section 3, our approach to tag recommendation based on the statistical language model theory will be presented and the expansion methods are also introduced. In Section 4, we setup the experimental platform. And then in Section 5, the effectiveness and efficiency of our proposals in tag recommendation are empirically verified. Section 6 concludes the paper by summarizing our work and discussing the future directions.

2. Related work

In this section, we first review the latest advances within the research area of tag recommendation and then survey some methods on language model.

2.1. Tag recommendation

Tag recommendation problem can be majorly divided into two application domains.

One domain is aimed at recommending tags for media resources such as pictures, audios and videos. Researches (Ames & Naaman, 2007; Liu, Hua, Yang, Wang, & Zhang, 2009; Sigurbjörnsson & Van Zwol, 2008; Wu, Yang, Yu, & Hua, 2009) focusing on this domain are majorly concerning on online expanding the existing tag set from users to encourage them post more tags, their methods are mostly inherits from the researches on query expansion or keyword expansion.

Another application domain concerns on recommending tags for documents (Brooks & Montanez, 2006; Golder & Huberman, 2006; Heymann, Ramage, & Garcia-Molina, 2008; Mishne, 2006; Song et al., 2008; Sood, Owsley, Hammond, & Birnbaum, 2007; Xu, Fu, Mao, & Su, 2006; Yan, & Hauprmann, 2007), and our work is focusing on this domain either. Brooks and Montanez (2006) developed a system, which can automatically tag blog documents based on the top three terms extracted from the documents, using TFIDF scoring. Their method inherits from a similar research domain called keywords extraction (Frank, Paynter, Witten, Gutwin, & Nevill-Manning, 1999; Turney, 2000), which has been studied for years. However, the keyword extraction problem is viewed as a subset of keywords generation, because it only extracts the keywords/keyphrases from the content of document but ignores those tags from more abstract level which do not appear in the document content. Chirita, Costache, Nejdl, and Handschuh, (2007) proposed more deeper methods by producing tag from both document content and the data residing on the user's Desktop, which could somehow overcome the drawbacks of keywords extraction, but its application environment is quite limited, because it relies on the personal data which is not easy to be obtained. These approaches are also known as Text Mining based approach.

Collaborative-filtering-based method is another popular scenario for tag recommendation. Mishne (2006) proposed a simple collaborative-filtering-based tagging system called "*AutoTag*", which finds the similar tagged documents and suggests some set of the associated tags to a user for selection. Sood et al. (2007) improved this thought by introducing tag compression and case evaluation to filter and rank tag suggestions. In opposites to Text Mining based approaches, tags recommended by collaborative-based approaches are mainly concerns on the high abstraction levels, because tags are aggregated from those already tagged documents, and those high abstraction tags are more common among these documents, and easier to be pushed out.

Most similar to our work, Song et al. (2008), proposed a clustering and classification based tag recommendation system, which partitions tags with documents into different clusters, and then classifies new documents into those clusters with a two-way Poisson Mixture Model, and tags belongs to the certain cluster are recommended, it can be viewed as a multi-label text classification based approach, and it overcomes both problems from Text Mining-based approach and collaborative-based approach. In this paper, we simplify their idea by treating each tag as a cluster, and further associating the new document directly to the tag, rather than a group of clustered tags which cannot easily calculates the similarity between document and each tag among the tags cluster. Also, we consider the tag recommendation problem is more like a ranking problem rather than the classification problem, because there are no fixed rich classes, but dynamic open tags which do not contains fixed documents for partitioning.

2.2. Language modeling

The statistical language model could assign a probability to a sequence of m words by means of a probability distribution, as $P(w_1, w_2, \ldots, w_m)$ or $P(w_1, m)$. Estimating the probabilities of word sequence may be expensive, since the sentence can be long and the size of corpus must be extremely large to avoid the data sparseness problem. In practice, the statistical language model is often approximated by smoothed n-gram models based on the Markov property, and the probability $P(w_1, m)$ can be represented as,

$$P(w_1, m) = \prod_{i=1}^{m} P(w_i | w_1, w_2, \ldots, w_{i-1}) \approx \prod_{i=1}^{m} P(w_i | w_{i-(n-1)}, \ldots, w_{i-2}, w_{i-1}) \quad (1)$$

Given different n, there are corresponding n-gram models, and the most common used models are,

$$\text{Unigram}: \quad P(w_1, m) = \prod_{i=1}^{m} P(w_i) \quad (2)$$

$$\text{Bigram}: \quad P(w_1, m) =) = \prod_{i=1}^{m} P(w_i | w_{i-1}) \quad (3)$$

Language modeling approach has been successfully introduced by Ponte and Croft (1998) in information retrieval. Given a document and a query, it treats the similarity between document and query as the probability of generating query from the language model formed by the document, known as LMIR. LMIR has been studied