Information Sciences 280 (2014) 16–25

Contents lists available at ScienceDirect

Information Sciences

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

Topic sense induction from social tags based on non-negative matrix factorization

Junpeng Chen [a,b,1], Shuai Feng [a,1], Juan Liu [a,*]

[a] School of Computer, Wuhan University, Wuhan 430072, PR China
[b] School of Information Engineering, Nanjing University of Finance & Economic, PR China

ARTICLE INFO

Article history:
Received 19 August 2013
Received in revised form 14 April 2014
Accepted 27 April 2014
Available online 6 May 2014

Keywords:
Topic sense induction
Social tag
Disambiguation
Non-negative matrix factorization

ABSTRACT

Social tagging, also noted as collaborative tagging or folksonomy, is an important way for users themselves to describe resources on the Web. The tags that the web users adopt to describe the resources are called social tags, and they have been widely used and studied. However, for the absence of a central controlled vocabulary, the semantics of the social tags are ambiguous due to constant changes of either the users' interests or the informal definitions, which makes it hard to directly make use of these social tags in the web applications. In this paper, we propose a non-negative matrix factorization (NMF) based method to automatically induce topic senses from social tags, which can then be used for the tag disambiguation. A novel automatic evaluation method is also proposed to evaluate our method. The experiment results show that the proposed topic sense induction method can help to provide precise resources search and recommendation, which is one of the key functionalities in social tagging systems.

© 2014 Elsevier Inc. All rights reserved.

1. Introduction

As one of the characteristics of Web 2.0 services, social tagging, also noted as collaborative tagging or folksonomy, allows users to collectively classify and annotate online resources. A large number of annotations taking together form a complex network of user-tag-resource triplets, commonly referred as a social tagging system. Since social tagging systems became popular around 2004, many social websites, such as Delicious, Flickr, and Last.fm have provided web services based on collaborative tagging, by which users can tag, search and browse resources from the web. It has been shown that social tagging systems benefit for organizational knowledge creating and resource sharing [5,33,35]. In addition, social tagging may hold the key for intelligent web services, semantic webs, and so on [3,14,17,23,28,31]. As valuable sources of knowledge resources, the folksonomies can be widely used in many communities such as information retrieval, ontology, and so forth. In fact, there are already a growth of the researches and applications of social tags in different domains [4,8,13,32,34,36,38]. Furthermore, some libraries have adopted social tags in catalog search and book recommendation [29,37].

Though social tags are very useful, there are several obstacles that may hinder the applications of social tags in knowledge sharing and intelligent service, due to the free-form nature of tagging and the lack of explicit semantics in social tagging systems:

* Corresponding author. Tel.: +86 2768775711.
E-mail address: liujuan@whu.edu.cn (J. Liu).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.ins.2014.04.048
0020-0255/© 2014 Elsevier Inc. All rights reserved.

(1) Syntactic variations

The same word in different syntactic forms may be used in various tags (terms). For instance, a user may annotate a web resource with the tag "picture", while another user may annotate the same resource with "pictures". These variations must be considered in intelligent web service in order to provide satisfied performance, or else, they may lead to confusion in web search and resource discovery [2].

(2) Polysemy

One tag may have multiple senses. For example, when a user uses the tag "java" to annotate a web resource, he may mean "Java island", or "java programming language".

(3) Synonym

Multiple tags may have the same sense. When a photograph is annotated, for instance, it may be annotated by one user as a "picture", while another user may think it is a "image". If a user uses either of them as the query keyword to a search engine, s/he may miss some important information because the related resources may be annotated by another different tag with the same sense (synonyms).

(4) Hypernym/Hyponym

The meaning of a tag may contain or be contained by other tags. When a user wants to annotate a new published video game, s/he may tag it as a "game" or a "video game". Obviously, the latter concept is more particular and will be more useful for recommendation systems to provide advice to those who want to find the interested video games.

To tackle the problems mentioned above, one can make use of word sense disambiguation (WSD) to assign sense labels to tags in social tagging systems. Most of the WSD methods rely on a pre-defined sense inventory that links terms to the corresponding concepts, so as to find the most similar sense for the target tag through the similarities [2,18,20,21]. In social tagging systems, however, such methods may be not applicable due to the following reasons:

(1) New words

New words appear every day, resulting that the pre-defined sense inventory is hard to cover them efficiently. For example, people may not find the meaning of the tag "iphone 5s" in a dictionary published two years ago.

(2) New meanings

The old words may have new meanings which cannot be found in a pre-defined sense inventory. For the word "python", it is referred to a kind of programming language just before recently.

(3) Limited coverage

Different social tagging systems may belong to different domains, and it is hard for a pre-defined sense inventory to cover all such domains. As shown in [2], WordNet, a most widely used sense inventory in WSD, only covers less than half of the terms used by the users of the social tagging system. Even though there are methods that can extract word sense inventories from Wikipedia or DBpedia, they can cover only parts of the tags occurring in social tagging systems too.

(4) Lack of domain knowledge

Different social tagging systems focus on different domains, and the fixed sense inventory is always lack of the important domain knowledge for helping people to disambiguate the tags. For example, the tag "VC" often represents "Venture Capital" in the financial domain, whereas it represents "Visual C++" in the computer science domain.

Compared with WSD, Word Sense Induction (WSI) can automatically identify the senses of words in the context without the pre-built sense inventory, hence WSI should be more applicable to attach appropriate semantics to tags. However, most of the present WSI methods mainly focus on plain texts, which cannot be directly adapted in social tagging systems that are organized as the networks of user-tag-resource triplets.

In this work, we propose an automatic WSI method, by which the tag senses specific to a particular social tagging system can be exactly and automatically extracted. Correspondingly, the sense inventory straightforwardly adapted to the domain can be built. Now that there are usually multiple topics hidden in the tags, we regard the semantic meanings of the tags as topic senses describing different topic semantics in various domains. The topic senses of the same tag (word) in different domains could be discriminated according to the different meanings of the word in different domains. Since many tags (words) in the same domain may stand for the same topic sense, thus the topic sense can be regarded as a profile of words.

Download English Version:

https://daneshyari.com/en/article/391628

Download Persian Version:

https://daneshyari.com/article/391628

Daneshyari.com