Expert Systems with Applications 42 (2015) 642–648

Contents lists available at ScienceDirect

Expert Systems with Applications

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

A classification approach for less popular webpages based on latent semantic analysis and rough set model

Jun Wang [a,*], Jiaxu Peng [a,*], Ou Liu [b]

[a] School of Economics and Management, BeiHang University, Beijing 100191, PR China
[b] School of Accounting and Finance, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

ARTICLE INFO

Article history:
Available online 19 August 2014

Keywords:
Webpage classification
Complex network analysis
Rough set
Latent semantic analysis

ABSTRACT

Nowadays, with the explosive growth of web information, the webpage classification faces great challenge. Computers have difficulty in understanding the semantic meaning of textual or non-textual webpages. Fortunately, Web 2.0 based collaborative tagging system brings new opportunities to solve this problem. It abstracts structured tags from unstructured content in webpages. However, large numbers of webpages on the Internet are less popular. Their tagging information is sparse, which makes their topic unclear and leads to ambiguous classification. Illuminated by the "ambiguous classification", we name the less popular webpage "hesitant webpage". In this paper, we propose an advanced approach for hesitant webpages classification. Firstly, hesitant webpages are divided into bridges, hubs and attached webpages according to their roles on the Internet. Secondly, attached webpages are classified by mining and extending their information in two perspectives. One is the latent semantic analysis (LSA) which is applied to fully explore the semantic meaning of sparse tags. It promotes accurate cognition of webpages semantically close to attached webpages. Another is the proposed density-relation-based rough set model which measures the affiliation degree of attached webpages in different categories. Experiment on real data shows that our approach effectively classifies the hesitant webpages base on the semantic meaning.

1. Introduction

In the information explosion era, web directory greatly reduces the number of returned webpages pushed by keyword search. This task is realized by classifying the webpages based on the semantic content. However, it is faced with great challenge resulted from the tremendous amount of web information. Besides, the webpages usually consist of textual information, images, audios, etc. All the information is unstructured and sets obstacles for computers to catch the semantic themes of the webpages. One good method to handle the unstructured data is tagging. It abstracts structured tags from unstructured content in webpages. Nowadays, a number of prominent web sites feature collaborative tagging which allows users to tag and share content publicly (Golder & Huberman, 2012). Tagging has been regarded as a possible solution to improve the searching of networked resources, as well as a means to support the personalized use (Guy & Tonkin, 2006; Ñanculef, Flaounas, & Cristianini, 2014; Trant, 2009;).

However, the cold reality is that, a large number of infrequently requested websites exist on the Internet (Kumar, Norris, & Sun, 2009). These webpages get sparse tags, leading to the ambiguity of their classification. In this paper, the less popular and sparsely tagged webpages are called hesitant webpages. These webpages share some common points with the tail data (Anderson, 2006). Though the tail resources are less popular, they might begin to show their power (Anderson & Andersson, 2007). Under this circumstance, classifying hesitant webpages properly has profound meaning in promoting information utilization and retrieval in networks.

Traditional classification methods cannot deal with the hesitant webpages properly. Partition-based methods, like the *k*-means (MacQueen, 1967), and density-based methods, like the DBSCAN (Ester, Kriegel, Sander, & Xu, 1996), tend to treat hesitant webpages as noises or outliers. Recently, Huang et al. proposed DenShrink (Huang, Sun, Han, & Feng, 2011). It efficiently reveals the embedded hierarchical community structure (Hastie, Tibshirani, & Friedman, 2001) and identifies hubs and outliers. But we found that it simply treats hesitant webpages as hubs and ignores the classification of them.

* Corresponding authors.
E-mail addresses: king.wang@buaa.edu.cn (J. Wang), peng.jia.xu@gmail.com (J. Peng), afliuou@polyu.edu.hk (O. Liu).

http://dx.doi.org/10.1016/j.eswa.2014.08.013
0957-4174/

In this paper, a classification approach for hesitant webpages is proposed. Firstly, the hesitant webpages are refined into three concrete types: bridges, hubs and attached webpages. Bridges and hubs are the overlapping portions and junctions of different categories while attached webpages are later classified. Secondly, considering the scarce tags of attached webpages, the tacit information of attached webpages is fully excavated in two perspectives. One is applying the latent semantic analysis (LSA) (Deerwester, Dumais, Landauer, Furnas, & Harshman, 1990) to the tags of all the webpages. By doing this, webpages which are semantically close with the attached webpage can be more clearly identified. The other perspective is the density-relation-based rough set model. It is built to measure the affiliation degree of attached webpage in different categories. This measurement takes not only the attached webpage's neighbors but also the dense pairs (Huang et al., 2011) of the neighbors into consideration. Experiment of real data shows that these two measures help to find semantically related categories for hesitant webpages and also enhance a classification measurement, the similarity based modularity (Feng, Xu, Yuruk, & Schweiger, 2007).

2. Related work

In this section, we will introduce the relation between the LSA and the classification of webpages, DenShrink clustering method and its limitation, and the rough set theory.

2.1. Latent semantic analysis

In the tagging systems, a webpage numbered i can be denoted by its tag set using the vector $p_i = (t_{i1}, t_{i2}, \ldots, t_{im})$. Wherein, m is the number of tags appearing in all the webpages. t_{ij} presents that a webpage i has a tag numbered j and the tagging frequency is t_{ij}. Since there might be zero, one or more users tagging webpage i with tag j, the tagging frequency is a non-negative integer. We adopt the vector space model to represent all the webpages with the tag-webpage matrix P. Wherein, n is the number of webpages.

$$\mathrm{P} = \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ \vdots \\ p_n \end{bmatrix} = \begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1m} \\ t_{21} & t_{22} & \cdots & t_{2m} \\ \vdots & \vdots & & \\ t_{n1} & t_{n2} & \cdots & t_{nm} \end{bmatrix}$$

The key idea in latent semantic analysis (LSA) is to map high-dimensional vector space representation of text webpages, to a lower dimensional semantic space representation (Hofmann, 2001). This process can be realized through the singular value decomposition (SVD) on matrix P by retaining a certain number of the largest characteristic values and eliminating the noises of matrix P. With this method, the refactored P describes the semantic relation between the tags and the webpages more exactly. Let us denote the refactored P with P'. Then p'_i and p'_j in matrix P' are used to compute the cosine similarity between webpage i and j.

LSA is meaningful for hesitant webpage classification. It fully excavates of the sparse tag set of hesitant webpage by mining the semantic similarity between tags on a global scale. Let hesitant webpage *hp* be a film review webpage infrequently tagged with merely four tags – *Renaissance*, *Opera*, *Aesthetics* and *Legend Story*. With traditional vector analysis, it would be difficult to find webpages which get high similarity with *hp* because of the sparse tags and some of the rarely used tags like "*Renaissance*" and "*Aesthetics*". With the application of LSA to the tag sets of large amounts of webpages, we are able to find more tags semantically close to the existing sparse tag set, such as "*Drama*", "*Art*" and "*the 16th century && Europe*", similar to "*Opera*", "*Aesthetics*" and "*Renaissance*" respectively. By doing this, the tag set of *hp* is potentially extended and it is more possible to find semantically similar webpages.

2.2. DenShrink clustering method

DenShrink is proposed by Huang et al. The main process of DenShrink can be divided into two phases. First, the micro-cluster which consists of two or more densely connected nodes (or super-nodes) is detected. Then the micro-cluster whose mergence increases the similarity based modularity measure is merged and becomes a super-node. The two steps are repeated iteratively until no micro-cluster can merge to increase the modularity.

In DenShrink, all the remaining isolate nodes which have not been merged till the last iteration and connect with multiple clusters are considered as hubs. However, our essential analysis of DenShrink reveals that some of the isolated nodes are unqualified hubs. These webpages are remained isolate because of the low similarity with other webpages, rather than acting as the hub among different categories. If consider the centrality degree (Borgatti, 2005) and the predicted trust of the information distribution path (Chen, Lü, Shang, Zhang, & Zhou, 2012; Kim & Song, 2011; Nocera & Ursino, 2012) through them, they are unqualified hubs. In our research, they are named *hesitant webpages*. An advanced approach is proposed to classify the hesitant webpages.

2.3. Tolerance relation based rough set theory

The classical rough set theory, first proposed by Pawlak (1982), attracted great attention for its fundamental role in rule extraction and classification problems (Chu, Gao, Qiu, Li, & Shao, 2010; Kaya, Pınar, Erez, & Fidan, 2013). One of the key definitions of the classical rough set theory is the indiscernibility relation. Later, various extended rough set models relaxed the establishment conditions of the relations in different degree (Kryszkiewicz, 1998; Stefanowski & Tsoukiàs, 1999).

Kryszkiewicz (1998) proposed the tolerance relation based rough set model in incomplete information system. The tolerance relation is a binary relation. Rather than the classical indiscernibility relation which is an equivalence relation, it is reflexive and symmetric but not imposed to be transitive. For example, let X denote a set $\{x_1,x_2,x_3,x_4,x_5\}$. Wherein, $x_1 = 210$, $x_2 = 234$, $x_3 = 765$, $x_4 = 297$, $x_5 = 739$, and $\mathrm{R} = \{\langle x,y\rangle | \ x \in X \wedge y \in X \wedge x$ and y share one or more numbers$\}$. It can be seen that x_1 and x_2 share the number "2", which we denote as $x_1 \mathrm{R} x_2$. In a similar way, there is $x_2 \mathrm{R} x_3$ but not $x_1 \mathrm{R} x_3$. It is obvious that R is reflexive and symmetric but not transitive, thus R is a tolerance relation.

Based on the tolerance relation, the related definitions of tolerance relation based rough set model are as the following.

Definition 1 (*Tolerance Class*). Let X denote the domain of discourse and R be a tolerance relation. For any $x \in X$, $\mathrm{T}(x) = \{y \in X | x \mathrm{T} y\}$ is the tolerance class of x. Particularly, for any $x \in X$, there is $x \in \mathrm{T}(x)$.

$\mathrm{T}(x)$ presents the set of all objects in X whose relationship with x satisfy R. In the above example, the tolerance class of x_1 relating to relation R is $T(x_1) = \{x_1,x_2,x_4\}$.

Definition 2 (*Upper Approximation*). Let X denote the domain of discourse and R be a tolerance relation. For any subset $X' \subseteq X$, iff $\overline{X'} = \cup\{\mathrm{T}(x) | x \in X'\}$, $\overline{X'}$ is upper approximation of X'.

The meaning of upper approximation can be intuitively understood. With any element in set X' certainly belonging to X', any object in $\overline{X'}$ is possible belonging to X' since there exists a certain