Expert Systems with Applications 38 (2011) 12807–12817

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Social trend tracking by time series based social tagging clustering

Shihn-Yuarn Chen [a], Tzu-Ting Tseng [b], Hao-Ren Ke [c,*], Chuen-Tsai Sun [a]

[a] Department of Computer Science, National Chiao Tung University, No. 1001 Ta Hsueh Road., Hsinchu 300, Taiwan
[b] Institute of Information Management, National Chiao Tung University, No. 1001 Ta Hsueh Road., Hsinchu 300, Taiwan
[c] Graduate Institute of Library & Information Studies, National Taiwan Normal University, No. 162, He-ping East Road, Section 1, Taipei 10610, Taiwan

ARTICLE INFO

Keywords:
Web 2.0
Social tagging
Time series clustering
Event tracking

ABSTRACT

Social tagging is widely practiced in the Web 2.0 era. Users can annotate useful or interesting Web resources with keywords for future reference. Social tagging also facilitates sharing of Web resources. This study reviews the chronological variation of social tagging data and tracks social trends by clustering tag time series. The data corpus in this study is collected from Hemidemi.com. A tag is represented in a time series form according to its annotating Web pages. Then time series clustering is applied to group tag time series with similar patterns and trends in the same time period. Finally, the similarities between clusters in different time periods are calculated to determine which clusters have similar themes, and the trend variation of a specific tag in different time periods is also analyzed. The evaluation shows the recommendation accuracy of the proposed approach is about 75%. Besides, the case discussion also proves the proposed approach can track the social trends.

1. Introduction

Social tagging has recently become a widely used application on the Internet. This process involves bookmarking part or all of a website for future reference. Social tagging can be used at a variety of websites, such as online shopping systems like Amazon.com, photo sharing communities like Flickr.com, and bookmarking services like Delicious.com. When someone finds something interesting online, he/she can tag it with some keywords. Tagging is very similar to bookmarking the entire page, and is similarly accessible.

Tagging also allows users to collaborate with other people online, including sharing collections and tag navigating. By sharing collections, a user can understand what other users bookmark and how others describe the same resource by various tags. Different resources tagged with the same word may refer to different subject matter, and this phenomenon can be found by navigating resources through one tag. For example, the tag "world-series" may highlight news reports regarding the 2009 World Series between the New York Yankees and the Philadelphia Phillies, but may also tag news reports about 2008 World Series between the Philadelphia Phillies and the Tampa Bay Rays. Tags can also be used to track news events. For example, news about Barack Obama's career as a senator to his presidential campaign and inauguration can be tagged simply "Obama."

This study analyzes social tagging information on time line, and each tag is represented by its tagging resources. Time series clustering is then applied to group tags with similar theme and find out the trends of events. In our example, there are five tags: 奧運 (Olympic Games), 中國 (China), 北京 (Beijing), 政治 (Politics) and 台灣 (Taiwan). Table 1 lists the usages of these tags in five sequential time points: p1, p2, p3, p4 and p5. Ignoring the chronological factor, traditional clustering algorithms group 奧運 (Olympic Games) and 中國 (China) in the same cluster, because of their similar usage count. However, according to Fig. 1, which depicts the usages of the tags at timeline, it is observably that 中國 (China), 政治 (Politics) and 台灣 (Taiwan) have similar polyline trends. Similar trends indicate these three tags have more similar theme than 奧運 (Olympic Games) and 北京 (Beijing), and these three tags should be grouped in the same cluster.

This study applies time series clustering to find out tags with similar trends. Based on clustering results, users can find related tags and documents in a particular time period. In addition, related documents from different time periods can be retrieved by calculating the similarities between clusters in different time periods.

The rest of this paper is organized as follows. Section 2 reviews previous studies on social tagging, time series analysis and clustering algorithms. Section 3 describes the proposed approach, covering data pre-processing, time series representation, time series clustering, and recommendation. Section 4 evaluates and compares the proposed approach and the counterpart approach that does not take into account the chronological factor. Section 5 concludes with future proposals.

* Corresponding author. Tel.: +886 2 77345203; fax: +886 3 5718925.
E-mail addresses: sychen@cs.nctu.edu.tw (S.-Y. Chen), baberainy@gmail.com (T.-T. Tseng), clavenke@ntnu.edu.tw, claven@lib.nctu.edu.tw (H.-R. Ke), ctsun@cs.nctu.edu.tw (C.-T. Sun).

0957-4174/$ - see front matter ∅ 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.04.073

Table 1
Tag usage example.

	p_1	p_2	p_3	p_4	p_5	Total
奧運	40	20	0	2	0	62
中國	8	15	22	12	10	67
北京	10	11	0	6	8	35
政治	5	10	20	10	8	53
台灣	6	9	19	8	10	52

2. Related works

2.1. Social tagging and folksonomy

"Folksonomy" is derived from the words "folks" and "taxonomy." It means a classification created by ordinary people. Vander Wal defined the term folksonomy as, "*... the result of personal free tagging of information and objects for one's own retrieval. Tagging is performed in a social environment (shared and open). Act of tagging is done by the person consuming the information.*" (Vander Wal, 2005) Folksonomy also includes collaborative classification, collaborative tagging, free tagging, tagsonomy, etc. Folksonomy emphasizes the spirits of social classification, collaboratively creation, and typically flat name-spaces.

Folksonomy consists of three aspects: user, resource, and classification (Fig. 2) (Pu, 2007). The user aspect involves social and collaborative concepts; the Resource aspect involves media information; the classification aspect defines the classification rules.

Social tagging is one type of folksonomy. Users can use tags, which are indicative keywords to annotate, describe or classify useful information. Flickr and Delicious.com are examples of websites which promote social tagging. Flickr is a photo sharing website where pictures can be tagged, and Delicious.com is a bookmark service provider which allows user to tag bookmarked URLs. In these instances, users are both consumers and contributors of tags, and these tags can be used for classification, indexing, searching and browsing content.

2.2. Clustering algorithm

There are various clustering algorithms which can be divided into five categories (Han & Kamber, 2001): partitioning methods (e.g.: *k*-means and fuzzy *c*-means), hierarchical methods (e.g.: agglomerative and divisive hierarchical clustering), density-based methods (e.g.: DBSCAN), grid-based methods (e.g.: STING) and model-based methods (e.g.: SOM). Clustering algorithms usually only process static data. Among the various clustering algorithms, the partitioning methods are most commonly used. A partitioning clustering method usually has to determine the number of clusters in advance, and then reduces the value of a goal function by iterative clustering computations. The halting condition of a partitioning clustering method is usually a threshold value of the goal function or a specific iteration count. For example, the *k*-means algorithm clusters data into *k* groups, and its goal function is the sum of square error between the centroid of a cluster and data items in the cluster.

2.2.1. Hierarchical clustering

This study uses hierarchical clustering to group time series data; this subsection introduces hierarchical clustering in greater detail. There are two types of hierarchical clustering: agglomerative (Voorhees, 1986) and divisive (Hastie, Tibshirani, & Friedman, 2009). Fig. 3 illustrates an example of hierarchical clustering. Agglomerative hierarchical clustering initially represents each data item as a cluster, and iteratively merges the two closest clusters till the halting constraint is satisfied. Divisive hierarchical clustering is different from agglomerative. Divisive method groups all data items in one group at beginning, and splits a cluster into two most distant clusters iteratively till the halting constraint is reached.

The criteria to decide cluster merging or splitting is the distance between clusters. The four ways to measure the distance between two clusters are single linkage, complete linkage, average linkage and Ward's distance (Ward, 1963).

I. Single linkage: Fig. 4(a) illustrates single linkage distance measurement, which only considers the shortest distance between two clusters. The distance is $D(C_i, C_j) = min\ d(a,b)$, where a belongs to cluster C_i, and b belongs to cluster C_j.
II. Complete linkage: Fig. 4(b) shows complete linkage distance, which considers the longest distance between two clusters. The distance is $D(C_i, C_j) = max\ d(a,b)$, where a belongs to cluster C_i, and b belongs to cluster C_j.
III. Average linkage: Fig. 4(c) displays average linkage, which considers the average distance between all data item pairs across two clusters. The distance is $D(C_i, C_j) = (\Sigma d(a,b))/(|C_i||C_j|)$, where a belongs to cluster C_i, and b belongs to cluster C_j.
IV. Ward's distance: Fig. 4(d) depicts Ward's distance; it finds out the centroid of two clusters first, and then calculates the square sum of distances between all data items and the centroid. The distance is $D(C_i, C_j) = (\Sigma|a - m|^2)$, where a belongs to $C_i \cup C_j$, and m is the centroid of C_i andC_j.

In addition to distance measurement of clusters, hierarchical clustering also has to consider the halting constraint before executing. The halting constraint is usually the cluster count or the average distance between clusters.

2.3. Time series analysis

A time series is a sequence of successive data measured at uniform time intervals (Box & Jenkins, 1976). Time series data is a set

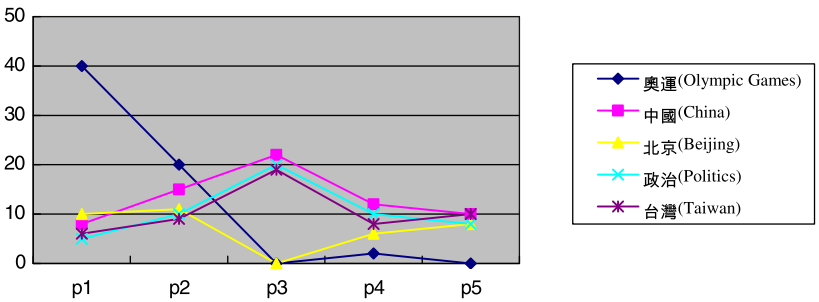

Fig. 1. Represent tags on time line.