Global Journal of Computer Science and Technology
Network, Web & Security
Volume 13 Issue 12 Version 1.0 Year 2013
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

An Enhanced Web Data Learning Method for Integrating Item, Tag and Value for Mining Web Contents

By R. Marutha Veni & P. Kavipriya
CMS College of Science and Commerce, India

Abstract - The Proposed System Analyses the scopes introduced by Web 2.0 and collaborative tagging systems, several challenges have to be addressed too, notably, the problem of information overload. Recommender systems are among the most successful approaches for increasing the level of relevant content over the "noise." Traditional recommender systems fail to address the requirements presented in collaborative tagging systems. This paper considers the problem of item recommendation in collaborative tagging systems. It is proposed to model data from collaborative tagging systems with three-mode tensors, in order to capture the three-way correlations between users, tags, and items. By applying multiway analysis, latent correlations are revealed, which help to improve the quality of recommendations. Moreover, a hybrid scheme is proposed that additionally considers content-based information that is extracted from items.

We propose an advanced data mining method using SVD that combines both tag and value similarity, item and user preference. SVD automatically extracts data from query result pages by first identifying and segmenting the query result records in the query result pages and then aligning the segmented query result records into a table, in which the data values from the same attribute are put into the same column. Specifically, we propose new techniques to handle the case when the query result records based on user preferences, which may be due to the presence of auxiliary information, such as a comment, recommendation or advertisement, and for handling any nested-structure that may exist in the query result records.

GJCST-E Classification : H.2.8

Strictly as per the compliance and regulations of:

An Enhanced Web Data Learning Method for Integrating Item, Tag and Value for Mining Web Contents

R. Marutha Veni [α] & P. Kavipriya [σ]

Abstract - The Proposed System Analyses the scopes introduced by Web 2.0 and collaborative tagging systems, several challenges have to be addressed too, notably, the problem of information overload. Recommender systems are among the most successful approaches for increasing the level of relevant content over the "noise." Traditional recommender systems fail to address the requirements presented in collaborative tagging systems. This paper considers the problem of item recommendation in collaborative tagging systems. It is proposed to model data from collaborative tagging systems with three-mode tensors, in order to capture the three-way correlations between users, tags, and items. By applying multiway analysis, latent correlations are revealed, which help to improve the quality of recommendations. Moreover, a hybrid scheme is proposed that additionally considers content-based information that is extracted from items.

We propose an advanced data mining method using SVD that combines both tag and value similarity, item and user preference. SVD automatically extracts data from query result pages by first identifying and segmenting the query result records in the query result pages and then aligning the segmented query result records into a table, in which the data values from the same attribute are put into the same column. Specifically, we propose new techniques to handle the case when the query result records based on user preferences, which may be due to the presence of auxiliary information, such as a comment, recommendation or advertisement, and for handling any nested-structure that may exist in the query result records.

I. Introduction

a) Data Mining

Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), is a field at the intersection of computer science and statistics, is the process that attempts to discover patterns in large data sets. It utilizes methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data preprocessing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.

The term is a buzzword, and is frequently misused to mean any form of large-scale data or information processing (collection, extraction, warehousing, analysis, and statistics) but is also generalized to any kind of computer decision support system, including artificial intelligence, machine learning, and business intelligence. In the proper use of the word, the key term is discovery, commonly defined as "detecting something new". Even the popular book "Data mining: Practical machine learning tools and techniques with Java" (which covers mostly machine learning material) was originally to be named just "Practical machine learning", and the term "data mining" was only added for marketing reasons. Often the more general terms "(large scale) data analysis", or "analytics" – or when referring to actual methods, artificial intelligence and machine learning – are more appropriate. According to one source, data mining is a marketing term coined by HNC, a San Diego-based company (now merged into FICO), at the beginning of the century to pitch their Data Mining Workstation.

The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining). This usually involves using database techniques such as spatial indexes. These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and predictive analytics. For example, the data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system. Neither the data collection, data preparation, nor result interpretation and reporting are part of the data mining step, but do belong to the overall KDD process as additional steps.

b) Web Mining

i. Web Structure Mining

Web structure mining is the process of using graph theory to analyze the node and connection structure of a web site. According to the type of web

Author α : MCA, Mphil, Asst. Professor, Dept of Computer Science, Dr. SNS RCAS, Coimbatore. E-mail : dhanuvene1@gmail.com
Author σ : MCA,(Mphil), Bharathiar University, Asst. Professor, School of commerce, CMS College of science and commerce, Coimbatore. E-mail : kavipriya.rajen@gmail.com

structural data, web structure mining can be divided into two kinds:

1. Extracting patterns from hyperlinks in the web: a hyperlink is a structural component that connects the web page to a different location.
2. Mining the document structure: analysis of the tree-like structure of page structures to describe HTML or XML tag usage.

ii. *Web Content Mining*

Mining, extraction and integration of useful data, information and knowledge from Web page contents. The heterogeneity and the lack of structure that permeates much of the ever expanding information sources on the World Wide Web, such as hypertext documents, makes automated discovery, organization, and search and indexing tools of the Internet and the World Wide Web such as Lycos, Alta Vista, WebCrawler, ALIWEB, MetaCrawler, and others provide some comfort to users, but they do not generally provide structural information nor categorize, filter, or interpret documents. In recent years these factors have prompted researchers to develop more intelligent tools for information retrieval, such as intelligent web agents, as well as to extend database and data mining techniques to provide a higher level of organization for semi-structured data available on the web. The agent-based approach to web mining involves the development of sophisticated AI systems that can act autonomously or semi-autonomously on behalf of a particular user, to discover and organize web-based information.

II. Literature Review

a) Mining Web Session Characteristic for Boundary Defense based on Hidden Markov Model

Yi Xie* and Xiangnong Huang

i. *Proposal*

Different from most existing studies on Web session identification, a novel dynamic real time user session processes description method is presented in this paper. The proposed scheme doesn't rely on presupposed threshold or client/server side data which are widely used in traditional session detection approaches. A new parameter is defined based on inter-arrival time of HTTP requests. A nonlinear algorithm is introduced for quantization. Nonparametric hidden semi-Markov model is applied to distinguish the user session processes. A probability function is derived for predicting user session processes. Experiments based on real HTTP traces of large-scale Web proxies are implemented to valid the proposal.

ii. *Drawbacks of the Proposal*

Different from traditional user session model, in this paper, a user session of a special user is divided into three segments: activity period, silent period and off-lining period. Activity period means the user is surfing the Internet, which causes the frequent interactions between user and different remote servers. Silent period indicates network connection is enable, but the user does nothing. During this period, the HTTP requests are mainly launched by softwares instead of user's own action. Thus, the number of requests in this period is far less than that of activity period. The last period means the network connection is unworkable or user has left.

b) Applying Concept Analysis to User- Session- Based Testing of Web Applications

i. *Proposal*

The continuous use of the Web for daily operations by businesses, consumers, and the government has created a great demand for reliable Web applications. One promising approach to testing the functionality of Web applications leverages the user session data collected by Web servers. User-session-based testing automatically generates test cases based on real user profiles. The key contribution of this paper is the application of concept analysis for clustering user sessions and a set of heuristics for test case selection. Existing incremental concept analysis algorithms are exploited to avoid collecting and maintaining large user-session data sets and to thus provide scalability. We have completely automated the process from user session collection and test suite reduction through test case replay. Our incremental test suite update algorithm, coupled with our experimental study, indicates that concept analysis provides a promising means for incrementally updating reduced test suites in response to newly captured user sessions with little loss in fault detection capability and program coverage.

ii. *Drawbacks of Proposal*

One approach to testing the functionality of Web applications that addresses the problems of the path based approaches is to utilize capture and replay mechanisms to record user-induced events, gather and convert them into scripts, and replay them for testing, Tools such as Web King and Rational Robot provide automated testing of Web applications by collecting data from users through minimal configuration changes to the Web server. The recorded events are typically base requests and name-value pairs (for example, form field data) sent as requests to the Web server. A base request for a Web application is the request type and resource location without the associated data. To our knowledge, these techniques do not include incremental approaches to test suite reduction.

c) Clustering and Tailoring User Session Data for Testing Web Applications

i. *Proposal*

Web applications have become major driving forces for world business. Effective and efficient testing of evolving web applications is essential for providing reliable services. In this paper, we present a user

session based testing technique that clusters user sessions based on the service profile and selects a set of representative user sessions from each cluster. Then each selected user session is tailored by augmentation with additional requests to cover the dependence relationships between web pages. The created test suite not only can significantly reduce the size of the collected user sessions, but is also viable to exercise fault sensitive paths. We conducted two empirical studies to investigate the effectiveness of our approach one was in a controlled environment using seeded faults, and the other was conducted on an industrial system with real faults. The results demonstrate that our approach consistently detected the majority of the known faults by using a relatively small number of test cases in both studies.

ii. *Drawbacks of the Proposal*

User-session based testing makes use of field data to create test cases, which has the great potential to efficiently generate test cases that can effectively detect residual faults. However, this approach is relatively new compared to traditional well developed techniques. There are several issues that must be addressed before it can serve as a sole testing method in practice. For an application that has been in production for a long time, the number of user sessions can be extremely large. Using all of the collected user session data requires much effort to determine which portion of the data can serve as the best representative of the system behavior. Nevertheless, a vast number of user sessions may not necessarily guarantee good coverage of the expected system behavior.

d) Separating Interleaved User Sessions from Web Log

i. *Proposal*

Analysis of user behavior on the Web presupposes a reliable reconstruction of the users' navigational activities. The quality of reconstructed sessions affects the result of Web usage mining. This paper presents a new approach for interleaved server session from Web server logs using m-order Markov model combined with a competitive algorithm. The proposed approach has the ability to reconstruct interleaved sessions from server logs. This capability makes our work distinct from other session reconstruction methods. The experiments show that our approach provides a significant improvement in regarding interleaved sessions compared to the traditional methods.

ii. *Drawbacks of the Proposal*

Session reconstruction is an essential data pre-process step in Web usage mining. The primary session reconstruction approaches based on time and reference cannot reconstruct the interleaved sessions and perform poorly when the client's IP address is not available. In this paper, an algorithm based on m-order Markov model is proposed which can reconstruct interleaved sessions from Web logs. The experiments show the promising result that the m-order Markov model has the ability to divide interleaved sessions, and can provide a further improvement when it is combined with the competitive approach.

e) Optimal Algorithms for Generation of User Session Sequences Using Server Side Web User Logs

i. *Proposal*

Identification of user session boundaries is one of the most important processes in the web usage mining for predictive prefetching of user next request based on their navigation behavior. This paper presents new techniques to identify user session boundaries by considering IPaddress, browsing agent, intersession and intrasession timeouts, immediate link analysis between referred pages and backward reference analysis without searching the whole tree representing the server pages. A complete set of user session sequences and the learning graph based on these user session sequences is also generated. Using this graph predictive prefetching is done. Comparison on the performance of the given approach with the existing reference length method and maximal reference method was done. Our analysis with different server's logs shows that our approach provides better results in terms of time complexity and precision to identify user session boundaries and also to generate all the relevant user session sequences.

ii. *Drawbacks of Proposal*

The analysis indicated that the existing web technology faces so many problems. One among them is personalization of web pages. Personalization is achieved if we know the browsing pattern of users. Our algorithm generates the efficient user session sequences with the less time complexity and good accuracy compared to the existing works. Thus we can reduce the latency. In the forth-coming papers USIDALG can be modified to generate the efficient learning graph to predict and prefect the user's next request.

iii. *The Existing Researches*

The Existing researches in personalizing the web user were single entity based and a summary of few researches are presented and the proposed system is developed by clearly understanding the below problems. This chapter discusses some of the existing techniques presented by different authors.

1. R. Cooley, B. Mobasher, and J. Srivastava, proposed a system for Navigation, in contrast to search, generally requires hierarchical storage, i.e. users need to create folders or directories and to store the information items "inside" them in preparation for future retrieval and use. Although other navigation methods have been proposed such as faceted classification and hypertext, neither is in

common usage in widely used operating systems, so we restrict our discussion here to common hierarchical methods. Hierarchical storage was first introduced to end-users in the *Multics* operating system in the mid 60s. Users were allocated a personal directory, in which they could create their own subdirectories, sub-subdirectories, etc., and store their files in any of these "locations." This directory structure was later applied in the Unix and the Linux operating systems.

2. O. Nasraoui, R. Krishnapuram, and A. Joshi, worked on the location metaphor became even clearer with the creation of digital folders first introduced in the *Xerox Star* in 1981. A folder is a visual metaphor for a location: users can see information items "inside" folders, as well as manipulate items and folders in various straightforward ways, e.g. drag and drop information items from one folder to another, etc.1 This folder hierarchy metaphor was later applied by Apple in the Mac operating systems and then by Microsoft in their Windows operating systems. Thus, location-based storage has been used without significant modifications, continuously and almost exclusively for several decades.
3. O. Nasraoui, R. Krishnapuram, H. Frigui, and A. Joshi, proposed a survey of unsupervised and semi-supervised clustering methods was presented by Grira, Crucianu and Boujemaa in. Squared error algorithms rely on the possibility of representing each cluster by a prototype. In general, the prototypes are the cluster centroids, as in the Kmeans algorithm. Fuzzy versions of methods based on the squared error are also defined, such as the Fuzzy C-Means. When compared to their 'crisp' counterparts, fuzzy methods are more successful in avoiding local minima of the cost function and can model situations where clusters actually overlap.
4. J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan In Morzy et al, Proposed a bottom-up approach of clustering based on Web Access Sequences is given, where frequent sequence patterns among web user sessions are identified. The users are then clustered based on their access sequence similarity. Shi has used the approach of fuzzy modelling taking into account the time duration that a user spends at a URL. Nasraoui et al have used the Competitive Agglomeration algorithm for Relational Data which yielded optimal number of clusters with non-Euclidean measures. In, it is argued that web user session identification itself is a non-trivial issue and clustering techniques have been used to characterise a user session. gives a basis of evaluating web usage mining approaches and for predicting the user's next request.
5. M. Spiliopoulou and L.C. Faulstich presented A survey of classification in data mining is given in. A sequence based clustering for web usage mining using K-means algorithm with artificial neural networks and Markov models is given in. It also demonstrates how a fuzzy approach yields superior accuracy. Artificial neural networks have been proven to be effective in dealing with classification problems and other machine learning areas. contains a brief tutorial of ANNs referred to in Section 6 and 7. Multilayered Perceptrons (MLP) were found to be appropriate for the dataset used. The applicability of MLPs is discussed. talks about Naive Bayes classifier which assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. Prefetching has been applied to a variety of distributed and parallel systems to hide communication latency
6. T. Yan, M. Jacobsen, H. Garcia-Molina, and U. Dayal, Crovella and Barford experimented and analyzed the effect of prefetching on the network performance by considering network delay as the primary cost factor. A simple transport rate controlled mechanism was proposed to improve the network performance. The usage of anchor text to index URL's in Google search engine was suggested by Brin and Page. The research focused on effective usage of additional information present in the hypertext. Chakrabarti et al designed and evaluated an automatic resource compilation system that could perform analysis of text and links to determine the web resources suitable for a particular topic.
7. M. Perkowitz and O. Etzioni Davison conducted a detailed analysis that focused on examining the descriptive quality of web pages and the presence of textual overlap in web pages. A keyword-based semantic prefetching approach was designed by Cheng and Ibrahim that could predict future requests based on semantic preferences of past retrieved Web documents. The scheme was evaluated by considering the Internet news services. Neural network was applied over the keyword set to predict future requests.
8. J. Borges and M. Levene, R.Cooley developed a quantitative model based on support logic that used information such as usage, content and structure to automatically identify interesting knowledge from web access patterns. The effectiveness of link-based and content-based ranking method in finding the web sites was analyzed by Craswell et al. The results indicated that anchor texts are highly useful in site finding.
9. O. Zaiane, M. Xin, and J. Han, Davison proposed a text analysis method that examined web page content to predict user's next request. The algorithm used text in and around the hypertext anchors of selected web pages to determine user's interest in

accessing web pages. Chen et al proposed a framework that used link analysis algorithm for exploiting both explicit (hyperlinks embedded in web page) and implicit (imagined by end-users) link structures. The framework had the ability to analyze interactions between users and the web.

10. O. Nasraoui and R. Krishnapuram developed a model including PageRank and HITS (Hypertext Induced Topic Selection) were the most popular webpage ranking algorithms. HITS emphasized on mutual reinforcement between the authority and hub web pages, whereas PageRank emphasized on hyperlink weight normalization. Ding et al generalized the concepts of mutual reinforcement and hyperlink weight normalization into a unified framework. Nadav and Kevin investigated anchor text to observe its relationship to titles, frequency of queries satisfied and the homogeneity of results obtained. Analysis indicated the anchor text resembled real-world queries in terms of its term distribution and length.
11. O. Nasraoui, C. Cardona, C. Rojas, and F. Gonzalez, Zhuge defined semantic links between resources to establish a high-level single semantic image to improve the quality of search result sets. The mathematical notations and formal structure of the semantic link network was presented in. Pierrakos et al clearly analyzed and presented the web usage mining process such as data collection, data preprocessing and pattern discovery that could be applied for web personalization. Jung carried out semantic outlier detection and segmentation using online web request streams to infer the relationships among web requests. A user support mechanism based on knowledge sharing with users through collaborative web browsing was proposed in [18]. It mainly focused on extracting user's interests from their own bookmarks.
12. P. Desikan and J. Srivastava, & Alexander Pons [19] proposed a technique that semantically bundled objects from slower loading web pages with objects of faster loading web pages. It was done to prefetch objects for the client's system prior to accessing the slower loading web page. carried out analysis on the web pages of different categories from Open Directory Project (ODP). They suggested the use of cohesive and non-cohesive text present near the anchor text to extract information about the target web page.
13. O. Nasraoui, C. Rojas, and C. Cardona researched a Access sequences as a criterion is not primary because these can be misleading in cases where the user does not know the ideal route to his destination. Also, considering sequences by themselves as a parameter has the risk of incorporating the undesirable step of giving equal importance to all sites, irrespective of the amount of time spent there, due to which the focus of the analysis is lost. In this paper, the time spent by a user at a URL is the criterion for analysing his degree of interest. The Naive Bayes Classifier is applied, following which, the K-means classification algorithm (statistical) is then compared with the Multilayer Perceptron (artificial neural networks) method using logged web usage data to analyse accuracy in classification.
14. C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, proposed Search as an Alternative to Navigation Through most of its long history, the hierarchical method has met with criticism. One disadvantage is that classification of information can 'hide' it from the user, and therefore reduce the chances of quick retrieval or reminding. In addition, the act of categorisation is itself cognitively challenging; users may find it hard to categorise information that could be stored in more than one category. Categorisation is also difficult because it requires that people anticipate future usage; moreover, that usage may change over time. At retrieval time, users need to recall how information was classified, which can be difficult when there are multiple categorisation possibilities. These problems were illustrated in a study of email categorisation. They found that users with many categories found it harder to file, and were more likely to create spurious unused folders. These apparent problems with navigation caused many PIM researchers and software developers to turn to *Search* as an alternative. There are intuitive potential advantages of search for both retrieval and organization. Search promises to be more flexible and efficient at *retrieval*, it does not depend on remembering the correct storage location; instead, users can specify in their query any attribute they happen to remember. They can also retrieve information via a single query instead of using multiple operations to laboriously navigate to the relevant part of their folder hierarchy. Regarding storage, search potentially finesses the *organizational problem* - as users don't have to engage in complex organizational strategies that exhaustively anticipate their future retrieval requirements. These arguments against navigation have been bolstered by recent developments in web access, where initial use of navigational systems such as Yahoo.
15. K. W. Church, W. Gale, P. Hanks, and D. Hindle, superseded a search engines such as Google The same logic has led to the development of experimental PIM search engines such as *Phlat*, *SIS*, *Haystack*, and *Raton Laveur*, as well as commercial systems such as *Einfish Personal*, *Copernic Desktop Search*, *Yahoo! Desktop Search* and *Microsoft Desktop Search*. Some more radical

systems such as *Lifestreams*, *Canon Cat, Presto*, *Placeless Documents*, *MyLifeBits*, and *Swiftware* explore alternatives to location-based hierarchies. However despite the rapid development of new such technologies, we know little about the effects of improved desktop search on user behaviour. In this study therefore we set out to test the following predictions about the effects of desktop search on both file retrieval and organization:

Retrieval: Search is more efficient and flexible for retrieval, thus improved quality of search engines should lead to a substantial increase in file search and eventually a preference for search over navigation.
File Organization: Users are known to have problems organizing files effectively for retrieval. Search allows retrieval without such manual organization and improved search should lead to a reduced use of filing strategies in preparation for later retrieval.

16. Q. Gan, J. Attenberg, A. Markowetz, and T. Suel worked on a Navigation or Search: Prior Evidence Pertaining to the Debate Evidence concerning users' *search* preference comes from empirical studies that examine retrieval behaviour. An early paper concerning users' retrieval habits, combined Barreau's interviews of novice personal computer users (using DOS, Windows 3.1 and OS/2) with Nardi's interviews of experienced Macintosh users. In both cases, users "overwhelmingly" preferred to navigate to their files than to search for them. Similar preferences for navigation were obtained in other more recent studies. These early findings raise a question– if search better suits users' requirements, why do they prefer navigation? One argument is that search technology is still immature. For example, Fertig and his colleagues argued that these navigation preferences result from limitations in search technology, and that improvements in search would inevitably lead to the replacement of navigation. They noted that the PIM search engines of that time (the mid 90s) were "slow, difficult, or only operate on file names (not content)" and did not provide incremental indexing. Fertig et al. further speculated that "inclusion of these better search techniques into current systems could sway results". However, their claim that the improvement of search engines would lead to an increased preference for search over navigation has not been tested empirically.
17. T. Joachims, presented few Other evidence challenging the effects of improved *search* concerns users' organizational efforts to prepare for future retrieval. There is some evidence that users seem to want to preserve folders, even when improved search is possible. Jones, Phuwanartnurak, Gill, & Bruce asked [14] participants the following question: "Suppose you could find your personal information using a simple search rather than your current folders.… Can we take your folders away?" Only one participant responded positively. In contrast, Dumais et al.'s participants tended to mildly agree with the sentence "I would likely to put less effort into maintaining a detailed set of folders for my files if I could depend on SIS (i.e., the *Stuff I've Seen* search engine) to find what I am looking for". Both studies asked whether the use of improved search engines would lead to less reliance on folders, but (perhaps because Jones et al. asked the question in a more extreme way) received different answers. Notice, however, that both researchers asked this as a hypothetical question.
18. K. W.-T. Leung, W. Ng, and D. L. Lee proposed lot of Improvements in Desktop Search Engines Today, more than a decade after Fertig et al.'s claims, commercial PIM search engines have improved considerably, newer search engines (such as *Google Desktop* and *Spotlight*) are better than the older ones (such as *Windows XP Search Companion* and Mac *Sherlock*) in the following ways:

Cross-format search: One limit of older search engines was that they allowed users to search only one format at a time. Following the SIS initiative, several improved search engines now support search across multiple datatypes – files, emails, instant messages and Web history within the same search query. This allows them to address the project fragmentation problem, where information items related to the same project but in different formats, are stored in different locations.

f) Existing System

The three types of recommendations in STSs (i.e., item, tag, and user recommendations) have been so far addressed separately by various approaches, which differ significantly to each other and have, in general, an ad hoc nature. Since in STSs all three types of recommendations are important, what is missing is a unified framework that can provide all recommendation types with a single method. Moreover, existing algorithms do not consider the three dimensions of the problem. In contrast, they split the threedimensional space into pair relations {user, item}, {user, tag}, and {tag, item}, that are two-dimensional, in order to apply already existing techniques like CF, link mining, etc. Therefore, they miss a part of the total interaction between the three dimensions. What is required is a method that is able to capture the three dimensions all together without reducing them into lower dimensions.

Finally, the existing approaches fail to reveal the latent associations between tags, users, and items. Latent associations exist due to three reasons:

1. Users have different interests for an item,
2. Items have multiple facets, and

3. Tags have different meanings for different users.

As an example, assume two users in an STSs for Web bookmarks (e.g., Del.icio.us, Bibsonomy). The first user is a car fan and tags a site about cars, whereas the other tags a site about wild cats. Both use the tag "jaguar." When they provide the tag "jaguar" to retrieve relevant sites, they will receive both sites (cars and wild cats). Therefore, what is required is a method that can discover the semantics that are carried by such latent associations, which in the previous example can help to understand the different meanings of the tag "jaguar."

III. Methodology

a) Singular Value Decomposition

Let X denote an m x n matrix of real-valued data and *rank* r, where without loss of generality $m \geq n$, and therefore $r \leq n$. In the case of microarray data, x_{ij} is the expression level of the i^{th} gene in the j^{th} assay. The elements of the i^{th} row of X form the n-dimensional vector $\boldsymbol{g}_i$, which we refer to as the *transcriptional response* of the i^{th} gene. Alternatively, the elements of the j^{th} column of X form the m-dimensional vector $\mathbf{a}_j$, which we refer to as the *expression profile* of the j^{th} assay.

The equation for singular value decomposition of X is the following:

$$X = USV^{\mathrm{T}} \tag{5.1}$$

Where U is an m x n matrix, S is an n x n diagonal matrix, and V^T is also an n x n matrix. The columns of U are called the *left singular vectors*, $\{\mathbf{u}_k\}$, and form an orthonormal basis for the assay expression profiles, so that $\mathbf{u}_i \cdot \mathbf{u}_j = 1$ for $i = j$, and $\mathbf{u}_i \cdot \mathbf{u}_j = 0$ otherwise. The rows of V^T contain the elements of the *right singular vectors*, $\{\mathbf{v}_k\}$, and form an orthonormal basis for the gene transcriptional responses. The elements of S are only nonzero on the diagonal, and are called the *singular values*. Thus, $S = \mathrm{diag}(s_1,\ldots,s_n)$. Furthermore, $s_k > 0$ for $1 \leq k \leq r$, and $s_i = 0$ for $(r+1) \leq k \leq n$. By convention, the ordering of the singular vectors is determined by high-to-low sorting of singular values, with the highest singular value in the upper left index of the S matrix. Note that for a square, symmetric matrix X, singular value decomposition is equivalent to diagonalization, or solution of the eigenvalue problem.

One important result of the SVD of X is that

$$X^{(l)} = \sum_{k=1}^{l} \mathbf{u}_k s_k \mathbf{v}_k^{\mathrm{T}} \tag{5.2}$$

is the closest rank-l matrix to X. The term "closest" means that $X^{(l)}$ minimizes the sum of the squares of the difference of the elements of X and $X^{(l)}$, $\sum_{ij} | x_{ij} - x^{(l)}_{ij} |^2$.

One way to calculate the SVD is to first calculate V^T and S by diagonalizing X^TX:

$$X^{\mathrm{T}} X = VS^2V^{\mathrm{T}} \tag{5.3}$$

and then to calculate U as follows:

$$U = XVS^{-1} \tag{5.4}$$

where the $(r+1),\ldots,n$ columns of V for which $s_k = 0$ are ignored in the matrix multiplication of Equation 5.4. Choices for the remaining n-r singular vectors in V or U may be calculated using the Gram-Schmidt orthogonalization process or some other extension method. In practice there are several methods for calculating the SVD that are of higher accuracy and speed. Section 4 lists some references on the mathematics and computation of SVD.

IV. Implementation and Findings

Given two data values f1 and f2 from different QRRs, we require their similarity, s12, to be a real value in [0, 1]. The data value similarity is calculated according to the data type tree shown in Fig. 4. Each child node is a subset of its parent node. For example, the "string" type includes several children data types, which are common on the Web such as "datetime", "float" and "price". The maximum depth of the data type tree is 4. In the following, we will refer to a non-string data type as a specific data type. Given two data values f1 and f2, we first judge their data types and then fit them as deeply as possible into the nodes n1 and n2 of the data type tree. For example, given a string "784", we will put it in node "integer".

The similarity s12 between two data values f1 and f2 with data type nodes n1 and n2 is defined as:

$$s_{12} = \begin{cases} 0.5 & n_1 = p(n_2) \,\&\, n_1 \neq String \quad OR \quad n_2 = p(n_1) \,\&\, n_2 \neq String \\ 1 & n_1 = n_2 \neq String \\ cosine\ \ similarity & n_1 = n_2 = String \\ 0 & otherwise \end{cases}$$

where p(ni) refers to the parent node of ni in the data type tree. The similarity between data values f1 and f2 is set to:

- 0.5, if they belong to different specific data types that have a common parent.
- 1, if they belong to the same specific data type.

- string cosine similarity of f1 and f2, if both f1 and f2 belong to the string data type.
- 0 otherwise, which occurs when one of f1 and f2 belongs to the string data type and the other one belongs to a specific data type, or f1 and f2 belong to different specific data types without any direct parent.

V. Conclusion & Future Enhancements

a) Conclusion

Social tagging systems provide recommendations to users based on what tags other users have used on items. In this paper, we developed a unified framework to model the three types of entities that exist in a social tagging system: users, items, and tags. We examined multiway analysis on data modeled as 3-order tensor, to reveal the latent semantic associations between users, items, and tags. The multiway latent semantic analysis and dimensionality reduction is performed by combining the HOSVD method with the Kernel-SVD smoothing technique. Our approach improves recommendations by capturing users multimodal perception of item/tag/user. Moreover, we study a problem of how to provide user recommendations, which can have significant applications in real systems but which have not been studied in depth so far in related research. We also performed experimental comparison of the proposed method against state-of the-art recommendations algorithms, with two real data sets (Last.fm and BibSonomy). Our results show significant improvements in terms of effectiveness measured through recall/precision. As future work, we intend to examine different methods for extending SVD to high-order tensors such as the Parallel Factor Analysis. We also indent to apply different weighting methods for the initial construction of a tensor. A different weighting policy for the tensor's initial values could improve the overall performance of our approach.

b) Future Enhancements

Although SVD has been shown to be an accurate data extraction method, it still suffers from some limitations. First, it requires at least two QRRs in the query result page. Second, any optional attribute that appears as the start node in a data region will be treated as auxiliary information. Third, similar to other related works, SVD mainly depends on tag structures to discover data values. Therefore, Finally, as previously mentioned, if a query result page has more than one data region that contains result records and the records in the different data regions are not similar to each other, then SVD will select only one of the data regions and discard the others.

References Références Referencias

1. E. Acar and B. Yener, "Unsupervised Multiway Data Analysis: A Literature Survey," IEEE Trans. Knowledge and Data Eng., vol. 21, no. 1, pp. 6-20, Jan. 2009.
2. N. Ali-Hasan and A. Adamic, "Expressing Social Relationships on the Blog through Links and Comments," Proc. Int'l Conf. Weblogs and Social Media (ICWSM), 2007.
3. M. Berry, S. Dumais, and G. O'Brien, "Using Linear Algebra for Intelligent Information Retrieval," SIAM Rev., vol. 37, no. 4, pp. 573-595, 1994.
4. M. Brand, "Incremental Singular Value Decomposition of Uncertain Data with Missing Values," Proc. European Conf. Computer Vision (ECCV '02), 2002.
5. J. Breese, D. Heckerman, and C. Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," Proc. Conf. Uncertainty in Artificial Intelligence, pp. 43-52, 1998.
6. D. Buttler, L. Liu, and C. Pu, "A Fully Automated Object Extraction System for the World Wide Web," Proc. 21st Int'l Conf. Distributed Computing Systems, pp. 361-370, 2001.
7. K. C.-C. Chang, B. He, C. Li, M. Patel, and Z. Zhang, "Structured Databases on the Web: Observations and Implications," SIGMOD Record, vol. 33, no. 3, pp. 61-70, 2004.
8. C.H. Chang and S.C. Lui, "IEPAD: Information Extraction Based on Pattern Discovery," Proc. 10th World Wide Web Conf., pp. 681-688, 2001.
9. L. Chen, H.M. Jamil, and N. Wang, "Automatic Composite Wrapper Generation for Semi-Structured Biological Data Based on Table Structure Identification," SIGMOD Record, vol. 33, no. 2, pp. 58-64, 2004.
10. W. Cohen, M. Hurst, and L. Jensen, "A Flexible Learning System for Wrapping Tables and Lists in HTML Documents," Proc. 11th World Wide Web Conf., pp. 232-241, 2002.
11. W. Cohen and L. Jensen, "A Structured Wrapper Induction System for Extracting Information from Semi-Structured Documents," Proc. IJCAI 2001 Workshop Adaptive Text Extraction and Mining, 2001.
12. V. Crescenzi, G. Mecca, and P. Merialdo, "Roadrunner: Towards Automatic Data Extraction from Large Web Sites," Proc. 27th Int'l Conf. Very Large Data Bases, pp. 109-118, 2001.
13. D.W. Embley, D.M. Campbell, Y.S. Jiang, S.W. Liddle, D.W. Lonsdale, Y.-K. Ng, and R.D. Smith, "Conceptual-model-based Data Extraction from Multiple-record Web Pages," Data and Knowledge Engineering, vol. 31, no. 3, pp. 227-251, 1999.
14. A. V. Goldberg, R. E. Tarjan. "A new approach to the maximum flow problem", Proceedings of the

eighteenth annual ACM symposium on Theory of computing, pp. 136–146, 1986.

15. D. Gusfield, Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology. Cambridge University Press, 1997.
16. B. Liu, W. S. Lee, P. S. Yu, and X. Li *proposed a Faster retrieval method for* Improved search engines are substantially faster than old ones. In some cases they have been demonstrated to be 1000 times faster.

User-centred design : Choosing between formats was not the only step the user had to take in older search engines. In addition, users had to choose between file name search or full text search, and also optionally specify the time the file was recently modified. To achieve a reasonable retrieval time, the user needed to input more information in order for the computer to do less, a feature which reflects a machine-oriented design. Newer search engines' retrieval speed allows them to reduce the query launching steps and complications to a minimum.

Incremental Search : One advantage of newer search engines is that they support incremental search, so that the search begins as soon as the user types the first character of the query. This has the benefit of being interactive: allowing users to refine their query in light of the results returned, and truncate the query after typing just a few characters if the target item is already in view. Older search engines were less efficient: prompting the user via form filling to specify multiple attribute fields and hit carriage return before the query is sent off. Incrementality, according to Raskin, has several advantages: (a) user and computer do not have to wait for each other, (b) users know they have typed enough to disambiguate their query because the desired file appears in the display, (c) users receive constant feedback as to the results of the search – they can correct spelling mistakes or refine search words without interrupting the search.

17. W. Ng, L. Deng, and D. L. Lee, proved that given these improvements in desktop search engines, it is now time to examine their implications: What are users' file retrieval preferences, what motivates retrieval by search, and what is the effect of improved desktop search engines on file retrieval preferences and file organization?

If the availability of these improved desktop search engines leads to a substantial increase in search, then it is reasonable to assume that this effect will continue to grow as search engines improve. If, on the other hand, no such effect is found, it raises questions regarding claims that improved search engines affect retrieval preferences and file organization, though it always can be claimed that future improvements in search could change this. As search engines are consistently improving and will continue to do so, the examination of their implications on PIM should be a continuous effort.

This page is intentionally left blank