Science Navigation Map: An Interactive Data Mining Tool for Literature Analysis

Yu Liu
School of Software
Dalian University of Technology
Dalian, China 116600
yuliu@dlut.edu.cn

Zhen Huang
School of Software
Dalian University of Technology
Dalian, China 116600
kobe_hz@163.com

Yizhou Yan
School of Software
Dalian University of Technology
Dalian, China 116600
yizhouyan9132@gmail.com

Yufeng Chen
School of Computer Science and Technology
Dalian University of Technology
Dalian, China 116600
cyxff1066@sina.com

ABSTRACT

With the advances of all research fields and web 2.0, scientific literature has been widely observed in digital libraries, citation databases, and social media. Its new properties, such as large volume, wide exhibition, and the complicated citation relationship in papers bring challenges to the management, analysis and exploring knowledge of scientific literature. In addition, although data mining techniques have been imported to scientific literature analysis tasks, they typically requires expert input and guidance, and returns static results to users after process, which makes them inflexible and not smart. Therefore, there is the need of a tool, which highly reflects article-level-metrics and combines human users and computer systems for analysis and exploring knowledge of scientific literature, as well as discovering and visualizing underlying interesting research topics. We design an online tool for literature navigation, filtering, and interactive data mining, named Science Navigation Map (SNM), which integrates information from online paper repositories, citation databases, etc. SNM provides visualization of article level metrics and interactive data mining which takes advantage of effective interaction between human users and computer systems to explore and extract knowledge from scientific literature and discover underlying interesting research topics. We also propose a multi-view non-negative matrix factorization and apply it to SNM as an interactive data mining tool, which can make better use of complicated multi-wise relationships in papers. In experiments, we visualize all the papers published at the journal of PLOS Biology from 2003 to 2012 in the navigation map and explore six relationship in papers for data mining. From this map, one can easily filter, analyse and explore knowledge of the papers through an interactive way.

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2741733.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*information filtering*

Keywords

Science Navigation Map, Interactive data mining, Multi-view non-negative matrix factorization

1. INTRODUCTION

Since the 21st century, all research fields and web 2.0 have developed rapidly, scientific literature increased dramatically online: digital libraries, citation databases, open access archives, etc.[4] In addition, due to abundant emerging metadata associated with papers and various scientific literature sources online, the relationships between papers become more and more complex. In general, nowadays, scientific literature has two main properties:

(a) Large volume, for example, there are more than 22 mililion papers for biomedical literature in PubMed and the number of the papers has increased to about 4% growth rate each year [9]. Over 1,153,000 papers were published in 2014, which means about 2.2 new biomedical papers were published by PubMed every minute.

(b) Complex multi-wise relations, various relationships between research papers can be identified from abundant literature-related data sources and text mining techniques. For example, direct link relation can be formed between papers according to citation or shared authors; while indirect relation can be inferred from text-based similarity, such as title and abstract similarity, or from simultaneous reference to or by third-party papers, such as co-reference [6] and co-citation [12].

On one hand, large volume of papers provide more abundant information to explore scientific literatures and various relations between papers give us different clues to comprehensively understand scientific context. They provide more potential reveal insight into research papers, which will help scientists to accelerate the discovery of new knowledge in scholarly communities, such as clustering and exploring internal structure of papers, finding research

topics, etc. On another hand, exploding volume of research papers and complex multi-wise relations between papers bring great challenges for scientists to accurately identify important and relevant research papers as well as explore and analyse papers.

To address these challenges, researchers have proposed many article level metrics (ALMs) based on online tools and Web 2.0 to filter papers [1, 8]. Many data mining techniques have also been imported to explore and analyse papers [10]. These methods have achieved certain effects. However, it is a problem to carefully apply ALMs to help research activities as well as data mining techniques typically require expert input and guidance, and return static results to users after processing, which makes them inflexible and not smart in practical applications [13, 2]. Therefore, we design an interactive data mining tool: Science Navigation Map (SNM) for literature navigation and analysis. It provides a graphical article level metric and supplies an interactive data mining tool where 'search', 'browse', 'filter', and 'topic discovery' can be carried out visually. When users search or browse related papers, SNM returns the paper lists as in traditional e-library, in addition all corresponding papers will also be highlighted on graphical navigation map at the same time. In this way, users can achieve global understanding of concerned paper distribution about impact, popularity and publication year simultaneously, and thus avoid losing directions in the sea of data. In addition, SNM provides the visualization of clustering results and keywords in each cluster. Clustering can be done with users' deep interaction, thus providing an integration of human cognition and data mining processes. In addition, we explore multi-wise relations between papers and propose a multi-view nonnegative matrix factorization (NMF), which can take advantage of the multi-wise relations to analyze papers. We apply multi-view NMF to SNM as a data mining technique for finding internal structure of papers and discovering underlying interesting research topics.

The remainder of the paper is organized as follows. In the following section, we explore multi-wise relations between papers and describe the proposed multi-view NMF. The third section introduces the experiment on our proposed interactive data mining tool: Science Navigation Map (SNM). Finally, we will present a conclusion.

2. METHOD

2.1 Paper Relationship exploration

Due to abundant emerging metadata associated with papers, the complex multi-wise relations between research papers give us different clues to comprehensively understand scientific context. These link relations can generally be classified as text-based similarity and bibliometric-based similarity. In Science Navigation Map (SNM), six link relations can be exploited, these adjacency matrices are shown in Figure 1. Two text-based similarity measures are computed based on title and abstract using the cosine similarity and the Term Frequency-Inverse Document Frequency model (TF-IDF) [11]. Cosine similarity is a similarity measure between two vectors. Given two feature vectors, A and B, the cosine similarity is represented by:

$$cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \quad (1)$$

Here, feature vectors of papers derived from all links types are nonnegative, so the cosine similarity of two papers ranges from 0 to 1. TF-IDF is widely used to represent the features of documents in information retrieval and text mining. A text document can be represented as a vector of index terms. A collection of documents can be represented by term-document matrix T where the weight of term i in document j can be calculated as:

$$T(i,j) = f_{ij} \log_2(N/N_i) \quad (2)$$

Where f_{ij} is the frequency of term i in document j; N is the total number of documents in the collection; N_i is the number of documents where term i appears. Depending on TF-IDF modeling, the term with a high term frequency and a low document frequency appears more informative. While the other four bibliometric-based similarity measures are derived from authorship, citation (one paper cites another paper), co-citation (another paper cites two paper simultaneously), and co-reference (two papers cite the same paper).

Therefore, six similarity matrices between papers are calculated based on title similarity, abstract similarity, author similarity, citation information, co-citation information and co-reference information as shown in Figure 1. Except for citation, the other five similarity matrices are calculated with the help of relation matrices derived from corresponding link types. These relation matrices work as feature matrices of papers that are employed to calculate the cosine similarity between papers. For title and abstract, the relational and term-paper matrices, are established with respect to the relation of papers to their terms; for author, author-paper matrix refers to the relation of papers and their authors; for co-citation matrix, citing paper-paper matrix refers to the relation of papers and their citing papers; for co-reference, reference-paper matrix refers to the relation of papers and their reference papers. The details of calculation are as the follows.

(a) The first view represents title similarity. $V^{(1)}(i,j)$ is the cosine similarity of the titles, which is calculated by means of the vectors of document i and j in term-paper matrix of titles. To avoid losing the information of papers with short titles, reduced stop word list is used.

(b) The second view represents abstract similarity. $V^{(2)}(i,j)$ is the cosine similarity of the abstracts, which is calculated by means of the vectors of document i and j in term-paper matrix of abstracts. In order to sparsity the slice, only scores greater than 0.2 (chosen heuristically to reduce the total number of non-zero elements, which is very effectively [3]) are retained.

(c) The third view represents author similarity. $V^{(3)}(i,j)$ is the cosine similarity of vectors i and j from author-paper matrix W where each element is computed as:

$$W(i,j) = \begin{cases} 1 & \text{if author } j \text{ wrote paper } i \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

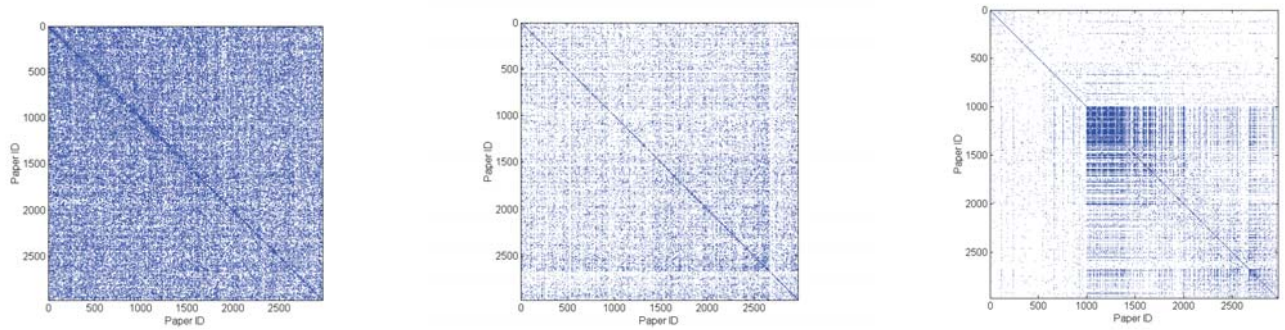
W is an $m \times n$ matrix where m is the number of papers; n is the number of authors (same name represents one author).

(d) The fourth view represents citation information, where each element is directly computed as:

$$V^{(4)}(i,j) = \begin{cases} 1 & \text{if paper } i \text{ cited paper } j \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

(e) The fifth view represents the co-citation information. $V^{(5)}(i,j)$ is the cosine similarity of the vectors i and j from citing paper-paper matrix W where each element is computed as:

$$W(i,j) = \begin{cases} 1 & \text{if paper } j \text{ cited paper } i \\ 0 & \text{otherwise} \end{cases} \quad (5)$$



(a) *Title* $V^{(1)}$ (b) *Abstract* $V^{(2)}$ (c) *Author* $V^{(3)}$

(d)*Citation* $V^{(4)}$ (e) *Cocitation* $V^{(5)}$ (f) *Coreference* $V^{(6)}$

Figure 1: The similarity matrices calculated from six perspectives: the title, the abstract, the author, the cited papers, the co-cited papers, and the papers co-referencing them

W is an $m \times n$ matrix where m is the number of papers; n is the total number of citing papers that may belong to any journal besides PLoS biology.

(f) The sixth view represents the co-reference information. $V^{(6)}(i,j)$ is the cosine similarity of the vectors i and j from reference paper-paper matrix W where each element is computed as:

$$W(i,j) = \begin{cases} 1 & \text{if paper } i \text{ refer to paper } j \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

W is a $m \times n$ matrix where m is the number of papers; n is the total number of reference papers that may belong to any journal besides PLoS biology.

2.2 Multi-view NMF

The relations between papers are represented as high dimensional matrices, non-negative matrix factorization (NMF) can be used to extract essential features and then give good low-rank approximations [7]. In contrast to principal component analysis (PCA) methods, NMF can achieve more intuitive physical interpretation of feature vectors by allowing only additive combinations of the non-negative basis vectors and sparse representation.

Given a non-negative data matrix $V \in R^{m \times n}$ and a positive integer $k \ll (m,n)$, NMF problems is to approximate V by computing a pair $W \in R^{m \times k}$ and $H \in R^{k \times n}$ to minimize the reconstruction error between V and WH. The objective can be mathematically formulated as the follows.

$$\min_{W \geq 0, H \geq 0} \|V - WH\|_F^2 \quad (7)$$

Where basis vector matrix W and coefficient matrix H are all non-negative. Each column of matrix W contains a basis vector while each column of H contains the weights needed to approximate the corresponding column in V using the basis from W. That is, each data vector v_i is approximated by a linear combination of the columns of W, weighted by the entries of i-th column of H, h_i. The greatest coefficient in h_i indicates the cluster this sample belongs to. Therefore, the clusters can be directly derived from the non-negative coefficient matrix H depending on a natural parts-based representation of NMF.

However, standard NMF just exploits an individual relationship at a time. It cannot deal with complicated multi-wise relationships in papers. Therefore, we propose a novel multi-view NMF, which exploits individual relationships in addition to integrating all link types to find interesting relationship. It regards different link types as different views to observe papers, which can integrate information from multiple views and take advantage of the latent information among different views.

The above six relations between papers can be regarded as six views used to observe the papers which are represented by six similarity matrices as shown in Figure 1: $V^{(1)}$, $V^{(2)}$, $V^{(3)}$, $V^{(4)}$, $V^{(5)}$, $V^{(6)}$. Then all the matrices are processed by NMF at the same time with some constraints. The objective function of multi-view NMF is as follows.

$$\min_{W^{(i)},H^{(i)}} (\sum_{i}^{n} \lambda^{(i)} \|V^{(i)} - W^{(i)}H^{(i)}\|_F^2 + \sum_{i \neq j}^{n} \lambda^{(i)}\lambda^{(j)} \|H^{(i)} - H^{(j)}\|_F^2)$$

$$s.t. \qquad i \in [1,n], W^{(i)} \geq 0, H^{(i)} \geq 0, \lambda^{(i)} \in [0,1], \sum_{i}^{n} \lambda^{(i)} = 1 \quad (8)$$

Where n is the number of views, here it is six. $W^{(i)} \in R^{m \times k}$ and $H^{(i)} \in R^{k \times n}$ are the corresponding basis matrix and coefficient matrix for i-th view. $\lambda^{(i)}$ is the weight of i-th view, which means the

Table 1: The information of each view

View(k)	Description	Non-zero	$\sum_i \sum_j V^{(k)}(i,j)$	View Weight
1	Title Similarity	127973	22585.2	5/32
2	Abstract Similarity	36631	6973.81	4/32
3	Author Similarity	39496	33715.62	1/32
4	Citation from PLoS Biology	319	638	20/32
5	CoCitation from all papers	9820	19114	1/32
6	CoReference	17609	29745	1/32

degree of influence on the model. To ensure the consensus clustering structure of all the $H^{(i)}$, all the views share the same reduced dimension k and extra constraint terms $\sum_{i\neq j}^{n} \lambda^{(i)}\lambda^{(j)}\|H^{(i)} - H^{(j)}\|_F^2$ are imported.

Equation (8) is not convex in $W^{(i)}$ and $H^{(i)}$ together. We refer to an algorithm of standard NMF for computation, that is, coordinate descent method [5]. It is to fix one variable while updating another variable alternately. According to coordinate descent method, we first update each element of $W^{(e)}$ to minimize the objective function while fixing all the matrices except $W^{(e)}$. For each element $W^{(e)}(r,t)$ of $W^{(e)}$, the update rules are formulated as the follows.

$$W^{(e)}(r,t) \leftarrow \max(0, W^{(e)}(r,t) - \frac{(W^{(e)}H^{(e)}H^{(e)T} - V^{(e)}H^{(e)T})(r,t)}{H^{(e)}H^{(e)T}(r,t)}) \tag{9}$$

Then, we can update each element of $H^{(e)}$ in the same way. The update rules for each element $H^{(e)}(r,t)$ of $H^{(e)}$ are as the follows.

$$H^{(e)}(r,t) \leftarrow \max(0, H^{(e)}(r,t) - \frac{(W^{(e)T}W^{(e)}H^{(e)} - W^{(e)T}V^{(e)})(r,t) + \sum_{i\neq e}^{n}\lambda^{(i)}(H^{(e)}(r,t) - H^{(i)}(r,t))}{W^{(e)T}W^{(e)}(r,t) + \sum_{i\neq e}^{n}\lambda^{(i)}}) \tag{10}$$

The matrices can be updated alternately following equations (9) and (10) until the model converges. Then serials of $H^{(i)}$ are obtained, which have the capacity of clustering papers according to standard NMF. In addition, because of the consistent clustering constraints imported in the model, all the $H^{(i)}$ have the same clustering structure. Therefore, all the $H^{(i)}$ can be combined to one clustering matrix as shown in equation (11):

$$H^{(final)} = \sum_{i=1}^{n} \lambda^{(i)} H^{(i)} \tag{11}$$

Where $\lambda^{(i)}$ is the corresponding weight of i-th view as in equation (8). $H^{(final)}$ can be used to cluster the papers as standard NMF.

From equations (8) and (11), we can find that if only one λ is set to 1 and other λs are set to 0, the multi-view NMF degenerates into standard NMF. It means that multi-view NMF can not only exploit the individual relation but also integrate all or even part of link types to find internal structure of papers by giving different sets. Its flexibility makes it a suitable tool for interactive data mining. Therefore, we exploit it in SNM.

3. EXPERIMENTS

3.1 Data Collection

We collected all the papers published in the journal of PloS Biology during 2003 to 2012 and obtained 2973 papers. The content of papers and the information of citation and social citation counts are collected from the online repository of PLoS biology, http://www.plosbiology.org/ and the open API provided by Public Library of Science (PLoS), http://api.plos.org respectively

Then we extract the collected data and calculate six similarity matrices between 2973 papers of PLoS biology as shown in Figure 1, where both x-axis and y-axis represent papers and they are sorted according to the order of publication dates. Each matrix displays the result of one similarity measure. If the paper i is connected to paper j, then $V(i,j)$ is non-zero and there will be a blue dot to represent this relationship. The statistic of similarity matrices are also listed in Table 1. Here 'Non-zero' denotes the ratio of non-zero elements against total elements in each matrix; The 'Sum' column is the sum of all non-zero elements. Here title similarity is denser than abstract similarity although abstracts include more words than titles. The reason is as follows. In abstract similarity matrix, only scores greater than 0.2 (chosen heuristically as [3]) are retained so that the number of non-zero elements are reduced. Reducing small elements makes representation of abstract similarity more effective. On the contrary, titles include few words originally, so the title similarity matrix is not reduced in order to preserve more information.

3.2 Data Mining via SNM

By using our propose multi-view NMF, we can achieve comprehensive relationships individually or simultaneously. Papers and relations with other paper can be showed in the SNM. SNM visualize the papers and provide a graphical article level metric as shown in Figure 2, where each paper is represented by a block, where size of the block denotes the paper impact (citation count) and color of the block denotes paper popularity (social citation count). When the cursor moves over the block, the corresponding paper information will show over the top of SNM. When a paper is selected, its related paper will also show on the SNM. Different link types can be specified.

The advantage of visual data exploration is that the user is directly involved in the data mining process, and then unifies the human intuition and cognition ability with advanced machine learning techniques to accelerate the discovery of new knowledge. SNM can facilitate interactivity in order to explore the latten research patterns and trends of data mining techniques within vast quantities of literature. First, the graphical article level metrics help to easily select high quality papers that interests them. Additionally, multi-view NMF is employed to find internal structure of the selected papers based on the paper relation matrices by giving view weights as shown in Table 1. Furthermore, the clusters of papers and top words in each cluster can be shown so that some hot research topics can be identified. New interesting topics may inspire the user to select other papers to analyze. Depending on such an iterative data mining process, the user can have a better understanding of the inherent structure involving a given set of papers and underlying interesting research topics.

The relations between papers are represented as high dimensional matrices, multi-view NMF can be used to extract essential fea-

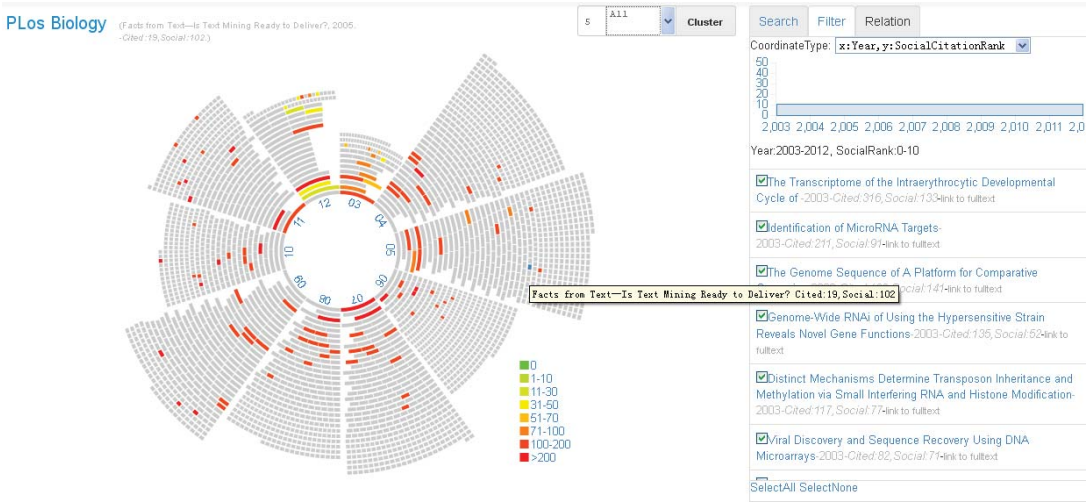


Figure 2: The 10 most popular papers of every year during 2003-2012.

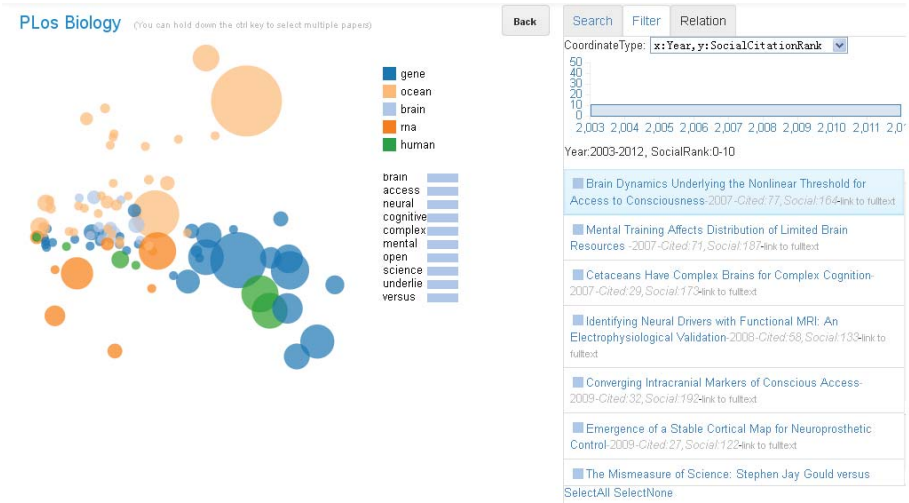


Figure 3: The cluster results: paper distribution with 2-dimensional view.

tures and then give good low-rank approximations. In contrast to P-CA methods, multi-view NMF can achieve more intuitive physical interpretation of feature vectors by allowing only additive combinations of the non-negative basis vectors and sparse representation.

The user can select concerned papers in SNM, and specify the number of clusters and link type for clusters, and then click the cluster button. The papers will be clustered depending on the multi-view NMF. The sparse representation of NMF can find good features in order to benefit cluster accuracy. However because of the sparse structure, all data points are near to coordinate axis, which is very suitable to computer handling rather than human observation. For easy graphical view of cluster results, PCA is employed to reveal paper distributions by reducing data to two dimensions. As shown in Figure 3, a bubble denotes a research paper whose color indicates the class it belong to and the size are proportionate to citation count. When the cursor moves on the bubble, the information of corresponding paper is shown on the top the SNM such as, title and publication year, impact(citation count), and popularity (social citation count). When the bubble is click, the papers and the other member papers belong to the same class are listed on the right of the SNM. When the cursor moves on the legend of a classes, the top keywords that appear in the abstracts of all papers of corresponding class are listed in decrease order of word frequency. Therefore, SNM helps researchers soon grasp important papers and hot topic in each cluster using multi-view NMF.

In a word, SNM provides intuitive and informative navigation and browsing, filtering, and knowledge discovery mechanisms for papers, papers relationships, and papers collection structure respectively. The user can analyze bibliometric data in a variety of interactive ways, so the flexibility, creativity, and general knowledge of the human combines with computational power of computers that facilitates and accelerates the discovery of new knowledge.

3.3 Discussion

Science Navigation Map (SNM) can be accessed through link http://www.linkscholar.org/plosbiology. All paper information, citations and social citations supporting SNM are from Public Library of Science (PLoS) which is the first publisher to provide an open application programming interface (API). This API allows developers to access article level metric data, such as usage statistics,

citation and social citations counts, blog and media, comments and ratings and gives researchers a huge opportunity to reveal insight into papers.

Similar to search engines and e-libraries, SNM also supplies the 'Search' and 'Related paper' functions. When users search or browse related papers, SNM returns the paper lists as in traditional e-library, in addition all corresponding papers will be highlighted on graphical navigation map at the same time. Therefore users can achieve global understanding of concerned paper distribution about impact, popularity, publication year of every paper simultaneously, and SNM helps users avoid losing directions in the sea of data.

Most distinctively, SNM supplies two innovative functions, 'Filter' and 'Clustering'. 'Filter' helps users to get high quality papers according to either the impact or the popularity of 2-dimensional metrics. The filtering criterions are 'CitationCount', 'CitationRank', 'SocialCitationCount', or 'SocialCitationRank' versus the years respectively. For example, when users wants to get the 10 most popular papers of each year during 2003 to 2012, they can select the option as 'x: Year, y: SocialCitationRank' for combo box 'CoordinateType' under 'Filter' tab on the right side of SNM, then draw a rectangle in the below coordinate system where the width is set to 2003-2012 and the height is set to 0-10. As shown in Figure 2, the filtering result is displayed on graphical navigation map with colorized blocks and paper lists sorted by years.

'Cluster' is a useful tool that helps researcher to cluster papers and finding heated topics. After selecting a set of papers, users must designate a number of clusters and standard of clusters from 7 link types, 'Abstract', 'Title', 'Author', 'Citation', 'Cocitation' and 'Coreference' and 'All'. In this example the most popular papers of 10 years are selected. The number of cluster is set to 5 and link type of 'All' is specified. The clustering results are visualized by colorful bubbles as illustrated in Figure 3. A bubble denotes a research paper whose color indicates the cluster it belongs to and the size is proportionate to citation counts. When the cursor moves on the bubble, information pertaining to the corresponding paper is shown on the top of the SNM such as, title and publication year, impact(citation count), and popularity (social citation count). When the bubble is clicked, the papers and the other member papers belong to the same cluster are listed on the right side. Each cluster is named by the keyword that ranks first in the cluster. The names of clusters accompanying corresponding colors are listed in capital letters on the right top of SNM. When the cursor moves on the name of a cluster, the top 10 keywords, which are included in the abstracts or titles (The two options can be appointed by users) of papers in corresponding cluster, are listed in decreasing order by Term weight whose options, TF (Term Frequency) or TF-IDF (Term Frequency-Inverse Document Frequency), can also be set by users according to their interests. Consequently, users can try different options to interactively explore research topics (keywords) underlying each cluster.

Through such iterative data mining, Science Navigation Map provides incredible opportunities to produce valuable insights of papers and inspire new ideas, thus is able to accelerate research.

4. CONCLUSION

In this paper, we explore multi-wise relations in papers and propose a multi-view NMF, which can take advantage of the multi-wise relations to analyze papers. We also apply the multi-view NMF to the proposed interactive data mining tool: Science Navigation Map (SNM) as a data mining technique. SNM provides a graphical article level metric and supplies an interactive data mining tool where 'search', 'browse', 'filter', and 'topic discovery' can be carried out visually. In the future, we will incorporate more advanced machine learning algorithms into SNM to enhance interactive data mining for literature so that users can easily handle massive and rapidly increasing literature, and then obtain more insights of papers and discovering interesting research topics and trends.

5. ACKNOWLEDGMENTS

This work was under Grand by the Natural Science Foundation of China and the Civil Aviation Administration of China jointly funded projects (No.U1233110), and the Fundamental Research Funds for the Central Universities (No. DUT13JR01).

6. REFERENCES

[1] E. Adie and W. Roe. Altmetric: enriching scholarly content with article-level discussion and metrics. *Learned Publishing*, 26(1):11–17, 2013.

[2] J. Demšar, B. Zupan, G. Leban, and T. Curk. *Orange: From experimental machine learning to interactive data mining*. Springer, 2004.

[3] D. M. Dunlavy, T. G. Kolda, and W. P. Kegelmeyer. Multilinear algebra for analyzing data with multiple linkages. *Graph Algorithms in the Language of Linear Algebra, J. Kepner and J. Gilbert, eds., Fundamentals of Algorithms, SIAM, Philadelphia*, pages 85–114, 2011.

[4] M. A. Haendel, N. A. Vasilevsky, and J. A. Wirz. Dealing with data: a case study on information and data management literacy. *PLoS biology*, 10(5):e1001339, 2012.

[5] C.-J. Hsieh and I. S. Dhillon. Fast coordinate descent methods with variable selection for non-negative matrix factorization. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1064–1072. ACM, 2011.

[6] M. M. Kessler. Bibliographic coupling between scientific papers. *American documentation*, 14(1):10–25, 1963.

[7] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[8] Y. Liu, Z. Huang, J. Fang, and Y. Yan. An article level metric in the context of research community. In *Proceedings of the companion publication of the 23rd international conference on World wide web companion*, pages 1197–1202. International World Wide Web Conferences Steering Committee, 2014.

[9] Z. Lu. Pubmed and beyond: a survey of web tools for searching biomedical literature. *Database*, 2011:baq036, 2011.

[10] V. G. Ribeiro, S. R. Silveira, A. da Silveira, R. Atkinson, and J. Zabadal. The use of data mining techniques for defining strategies in scientific communication processes in design journals. *Strategic Design Research Journal*, 6(2):85–94, 2014.

[11] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

[12] H. Small. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for information Science*, 24(4):265–269, 1973.

[13] V. Zelevinsky. Interactive data mining, May 31 2013. US Patent App. 13/906,540.