lysis of Lanczos-type methods for the linear response eigenvalue problem. *Journal of Computational and Applied Mathematics*, 247, p.p.17-33.

15. E. V. Tsiper (1999) Variational procedure and generalized Lanczos recursion for small-amplitude classical oscillations. *JETP Letter*, 70(11), p.p.751-755.
16. Z. Teng, Y. Zhou, and R.-C. Li (2013) A block Chebyshev-Davidson method for linear response eigenvalue problems. *Technical Report 2013-11, Department of Mathematics, University of Texas at Arlington*. Available at http://www.uta.edu/math/preprint/, submitted.
17. W. Kahan (1967) Inclusion theorems for clusters of eigenvalues of Hermitian matrices. *Technical report, Computer Science Department, University of Toronto*.
18. D. C. Dzeng and W. W. Lin (1991) Homotopy continuation method for the numerical solutions of generalised symmetric eigenvalue problems. *The Journal of the Australian Mathematical Society. Series B.* Applied Mathematics, 32(4), p.p.437-456.
19. J. Kovač-Striko , K. Veselić (1995) Trace minimization and definiteness of symmetric pencils. *Linear Algebra and its Applications*, 216, p.p.139-158.
20. X. Liang, R.-C. Li, and Z. Bai (2013) Trace minimization principles for positive semi-definite pencils. *Linear Algebra and its Applications*, 438, p.p.3085-3106,.
21. X. Liang and R.-C. Li. (2015) The hyperbolic quadratic eigenvalue problem. *Forum of Mathematics, Sigma*, p.p.93-93.

Semantic Related Feature Analysis and Dynamic Evolution Based on Topic Temporal Chains under the Social Network

Caiyin Wang

Intelligent Information Processing Laboratory, Suzhou University, Suzhou 234000, Anhui, China

Lin Cui[1, 2]

[1]Intelligent Information Processing Laboratory, Suzhou University, Suzhou 234000, Anhui, China
[2]College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics,
Nanjing 210016, Jiangsu, China

Xiaoyin Wu, Baosheng Yang

Intelligent Information Processing Laboratory, Suzhou University, Suzhou 234000, Anhui, China

Abstract
As traditional topic detection under the online social network is lack of mining the topic semantic features and dynamic evolution characteristics, this paper studies the semantic features and dynamic evolution trend based on topic temporal chain under the online social network. According to the temporal relations, text flows are divided into the text set in continuous time slice, then combination of cosine similarity and Jaccard similarity coefficient is used to implement semantic related feature analysis of topic temporal chain. Experiments on the two real social network are made to verify the topic evolution performance of the proposed method by comparison with the existed CW-TE(Co-word-based Topic Evolution) method and RE-TE (Relative-Entropy-based Topic Evolution) method, which demonstrate that the proposed online topic evolution algorithm can more effectively detect semantic related feature and dynamic evolution of topics and obtain better detection performance.

Key words: SOCIAL NETWORK, TOPIC TEMPORAL CHAIN, SEMANTIC RELEVANCE, DYNAMIC EVOLUTION

1. Introduction

With the deep research on the static properties of social network, researchers have begun to focus on the time properties of the social network [1]. How the time factor impacts the evolution of the social network becomes the new development direction of the social network. Traditional topic detection under the online social network mainly regards text topic extraction as the main goal, lacking the support on the semantic related features and dynamic evolution of topics [2]. The flow of online social network can be described as a dynamic network link structure and document contents changing with time, typical examples are that in micro-blog system, the relationship of concerning and being concerned among uses and posts always changes over time, which are the results of the interaction among users.

An important task of online social network analysis is how to discovery and track the changes of user interesting topics with time. There are a lot of research work on social network link structure flow and text flow, however, simultaneously considering both factors and their interaction are less. In this paper, comprehensively considering the topic temporal chain, network link structure and text semantic factors, the online topic-evolution model which is abbreviated as OTE is proposed. The proposed OTE model aims at the research of the topic semantic features and dynamic evolution trend of the online social network, fusing with the changes of the topic documents and network structure on the time axis, combining the cosine similarity and Jaccard Similarity Coefficient to realize semantic related feature analysis of topic timing chain, eventually establishing a probability model to complete topics evolution task under the online social network.

The remainder of this paper is organized as follows. Section 2 introduces the related work about the current topic evolution under the online social network. In section 3, some definitions used by this paper are provided in detail. Section 4 proposes the general framework of the proposed semantic related feature analysis and dynamic evolution method based on time temporal chains under the online social network. Section 5 makes some comparative experiments among two investigated algorithms with respect to precision rate, recall rate, F-measure value and their evolution trends, respectively. Finally, Section 6 draws some conclusions and points out some future work.

2. Related work

Online social network is a collection composed of large-scale users and their relatively stable relationship, which has become an important way for the daily communication of peoples [3]. To a certain extent, online social networks can be regarded as a kind of mapping that real social relations are in the cyberspace. Therefore, the online social network is a self-organization network which takes the user as the center, the topology structure and user information as the basic elements. The evolution mechanism of the network topology and the information communication mechanism from users are the inherent law of the online social network [4].

The social network is firstly originated from six degrees of separation theory and the rule of 150 in the 1960s [5]. In various important academic conferences and academic journals, there disappear issues and articles about online social network research. ACM, WWW, KDD and other related fields also opened up related topics [6]. After the sustainable development of some series of social media such as Friendster, MySpace and so on, Facebook has been widely used in 2004 [7], then in 2006,Twitter and other micro-blog websites have also been popular [8]. In China, Sina micro-blog produced far-reaching influence. Nowadays, online social network has become a global phenomenon, which has attracted a huge amount of users and constitute a real world network mapping.

The main researches of social network mainly include the discovery of key members and social communities, the extraction of feature pattern and evolution mode analysis and so on [9], among which, analysis of evolution model refers to that users produce large amounts of information in every day, update their status, create new social relations [10]. In the evolution process of social networks, some information can receive a lot of attention through the transmission and amplification in a very short period of time, but also affected by the small world effect, the information can be spread to the world quickly. The evolution of social networks analysis contains the topic model research, topic detection and tracking technology and topic clustering, etc [11]. Among them, the technology of topic detection and tracking was first proposed by American Defense Advanced Research Projects Agency, mainly studied how to find out important information from the mass of information in a timely manner, has become one of the research hot spots. At present, most of the research on topic detection and tracking only consider static data and closed data application background, ignoring the dynamic characteristics of topic data, cannot fully mine the variation characteristics of text information on the strength and the content of the topic evolution process [12].

In this paper, the characteristics variation of social networks is considered from a dynamic perspective, the formal description method of dynamic community and its individual are studied through analyzing the dynamic evolution process of social network, the dynamic behavior information of individual network and inter individual links, proposed a new dynamic attribute similarity between calculation method and the method of community recommendation algorithm, a calculation of individual and other individual behavior similarity using dynamic attribute similarity, the similarity of individual communities currently recommended to the individual. This method can be used as the individual through the similar behavior of individual recommendation community, there is no direct connection among similar indivi- duals.

3. The proposed Online Topic-Evolution Model (OTE)under the social network

Under online social network, the topic evolution can be divided into two aspects by using topic model, which are the change of topic content and the fluctuation of topic intensity in the time slice. Based on the study of online social networks in the topic detection, this paper firstly proposed the construction method of topic temporal relationship between the time chain and semantic relation extraction. The basic definitions on topic change and evolution required by this paper are defined as follows:

3.1. Theoretical foundation and hypothesis

In this paper, an online dynamic social networks is defined as a social network flow $G=\{N^t,E^t,P^t\}_{t=1}^{T}$, N^t represents the node set in t time in the social network, E^t denotes link set between nodes in t in the social network, P^t is the posts set or paper sets produced by nodes from *t-1* to *t*. In an online social network, a node denotes a user, an edge $e_{ij}{}^{t}\in E^T$ represents a link relation between user *i* and user *j* in time t, $e_{ij}{}^{t}$ is an arbitrary non negative value,When $e_{ij}{}^{t}$ is equal to 0, which indicates that there is no link relation between user *i* and user *j*. In addition, $P^t=\{p_i{}^{t}\}_{i=1}^{N}$, among which, $p_i{}^{t}$ is the post or the paper produced by user *i* from *t-1* to *t*,and $p_i{}^{t}$ is also a bag of words from the dictionary $W=\{w_v\}_{v=1}^{V}$.

Based on the definition of online social network, we can further formally define the topic evolution problem. Given a dynamic social network, the task is to establish a reasonable model to simulate the evolution of social networks. In order to better carry out modeling and tracking on the topic, the following independent intuitive hypotheses are considered that the topic intensity of the user at the time t is only dependent on the topic intensity of the users themselves in t-1 and the network structure. On the one hand, this assumption is based on the user tends to be influenced by the neighbors and following the interests of the neighbors, on the other hand, the interests of users are also influenced by their historical interests. The state of the network structure in the t time is only dependent on the topic intensity of users in the current time.

3.2. Framework design model

The framework of semantic related feature analysis and evolution analysis based on topic temporal chain is shown in Figure 1.

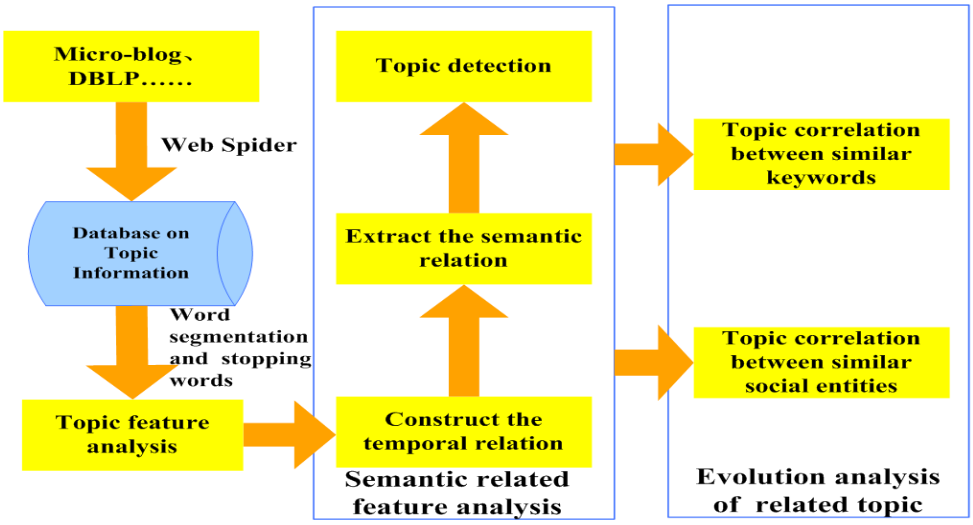
Micro-blog、DBLP……
Web Spider
Database on Topic Information
Word segmentation and stopping words
Topic feature analysis
Topic detection
Extract the semantic relation
Construct the temporal relation
Semantic related feature analysis
Topic correlation between similar keywords
Topic correlation between similar social entities
Evolution analysis of related topic

Figure 1. Framework of semantic related feature and evolution analysis based on topic temporal chain

As Figure 1 shows, semantic related feature and evolution analysis based on topic temporal chain mainly include three steps that are initialization pre-processing, topic detection and evolution analysis of related topics. Firstly, during the initialization pre-processing, the real datasets from Sina Micro-blog and DBLP are extracted through using web spider, then databases on topic information are got. Through word segmentation and stopping words, topic feature analysis is obtained. Secondly, topic detection is implemented, which contains constructing the temporal relation, extracting the semantic relation and executes topic detection. Lastly, Evolution analysis of related topic is facilitated, which includes topic correlation between similar keywords and between similar social entities.

3.3. Semantic related feature analysis of topic temporal chains

During the initialization process, English word segmentation tool and the Chinese word segmentation tool ICTCLAS are used to pre-process the word segmentation operation on the topic of the original corpus, respectively. Part of speech tagging and stopping words list are also used to remove the stopping words and duplicate features, etc.. Then the topic texts are represented as feature vectors of feature space, which are used to to extract the keywords set and adopt the TF-IDF method as the keyword weighting. TF-IDF computing method is as follows:

$$w_i(p) = TF - IDF_i = \frac{freq_i(p)}{\max_i\{freq_i(p)\}} \times \log\frac{N}{n_i} \qquad (1)$$

Among which, $freq_i(p)$ is the occurrence frequency of key words i in the post p, N is the number of topic texts, n_i is the number of keywords i.

After the topic texts are represented as feature vectors, in order to avoid the shortcomings of the cosine similarity, the similarity between the topic texts is calculated by using the combination of the cosine similarity and the Jaccard coefficient similarity. The similarity between feature vectors is measured through using the cosine similarity, the topic correlation of similar keywords is expressed as the formula (2).

$$sim(t_i,t_j) = \frac{\sum_{k=1}^{m} tw_k(t_i)\times tw_k(t_j)}{\sqrt{\sum_{k=1}^{m} w_k{}^2(t_i)}\times\sqrt{\sum_{k=1}^{m} w_k{}^2(t_j)}} \qquad (2)$$

Among which, $tw_i(t_1)$ and $tw_i(t_2)$ are respectively the *ith* keyword weights of the topic t_1 and t_2, $w_i(t_1)$ and $w_i(t_2)$ are respectively the weights where the *ith* keywords of the post t_1 Jaccard coefficient is used to calculate the similarity between similar social entities, the calculation formula is as follows:

$$sim_{jac}(t_i,t_j) = \frac{|t_i \cap t_j|}{|t_i \cup t_j|} \qquad (3)$$

Where, $|e_i \cap e_j|$ denotes the number of the public feature in the social entity i and j, $e_i \cup e_j$ represents the number of different features in the social entity i and j.

The weighted similarity combining cosine similarity with Jaccard coefficient is obtained, which not only consider the topic correlation of similar keywords, but also take into account the topic correlation between the similar social entities. The calculation method is shown in the formula (4):

$$sim(t_i,t_j) = sim_{\cos}(t_i,t_j)\times\alpha + sim_{jac}(t_i,t_j)\times(1-\alpha) \qquad (4)$$

Where, α denotes the contribution of cosine similarity to total similarity, $1-\alpha$ is the size of the contribution of Jaccard coefficient similarity to the total similarity.

3.4. The proposed OTE algorithm

Combined with Section 3.1,3.2 and 3.3, the proposed algorithm is as follows:

Algorithm 1: Semantic related feature analysis and dynamic evolution algorithm based on topic temporal chains under the social network

```
Input: $G = \{N^t, E^t, P^t\}_{t=1}^{T}$
Output:TopicEvolutionItem[ ]

1. For a given node x, initialization topic sets is defined as $S_x(0) = x$ and set the number of iterations t=1;
2. Preprocess(T) → topicTextArray[ ] // Feature vectors are obtained by using TI-IDF
3. for(int i=0;i<topicTextArray.length();i++)
4.  for(int j=i+1;j<topicTextArray.length();j++)
5.  {
6.  ClusterItem[i].clusterId =i;
7.  ClusterItem[j].clusterId =j;
8.  ClusterItem.next=null;
9.  Calculate the maximum sim(ClusterItem[i], ClusterIterm[j]) using the formular (4)
10.  Execute the clustering for ClusterItem by using K-means algorithm
11.  if( sim(ClusterItem[i], ClusterIterm[j]) > the specified threshold)
12.  make ClusterItem[j].clusterId = ClusterItem[i].clusterId
13.  else
14.  break;
15.  }
16.  TopicEvolutionAnalysis (t,x)
17.  {
18.  For any $x \in ClusterItem$ ,let $S_x(t) = f(S_{x_{i_1}}(t), \cdots S_{x_{i_m}}(t), S_{x_{i_{(m+1)}}}(t-1), \cdots, S_{x_{i_k}}(t-1)$ .
19.  If the label of each node keep consistent with tags of their most neighbor nodes, terminate the program;or t=t+1 and return to (18)
20.  }
```

4. Experimental result analysis

4.1. Experimental data sets setting and experimental conditions

In order to verify the proposed algorithm in this paper, we collected two experimental data sets. An experimental data set came from Sina micro-blog about ALS Ice Bucket Challenge (http://www.sina.com.cn/), another experimental data set came from DBLP about big data (http://dblp.uni-trier.de/db/). The detailed information on this two datasets is shown in Table 1:

Table 1. Distribution of datasets from Sina micro-blog and DBLP

Attributes	Sina micro-blog	DBLP
	ALS Ice Bucket Challenge	Big Data
The number of posts or papers	121,104	3,250

All the experiments were performed on a personal computer with an Intel Core2 Duo, 2.66-GHz CPU and 2 GB RAM. The operating system is Microsoft Windows 7. All the programs were coded in Java language.

4.2. Metrics and baselines

In order to verify the performance of the proposed OTE algorithm, precision rate, recall rate and F-measure value are used to evaluate its effectiveness in this paper. The specific indicators of the evaluation are based on the contingency matrix, which represents the number of documents that satisfy the requirements of the related matrix, as shown in the following Table 2.

Table 2. Correlation matrix of topic detection results

	In the labeled topic	Not in the labeled topic
In the detection topic results	a	b
Not in the detection topic results	c	d

Based on the correlation matrix, precision rate, recall rate, and F-measure value are defined, respectively:

$$precision = a/(a+b) \qquad ((a+b)>0, or \text{ precision is not defined}) \tag{5}$$

$$recall = a/(a+c) \qquad ((a+c)>0, or \text{ recall is not defined}) \tag{6}$$

$$F-measure = \frac{2 * precision * recall}{precision + recall} \tag{7}$$

In this paper, the following state-of-the-art methods are evaluated for the comparison, including our proposed OTE model, CW-TE model (Co-word-based Topic Evolution) [13]and RE-TE model (Relative-Entropy-based Topic Evolution) [14]. CW-TE model put up a feature selection approach for social media information in different time spans and then constructed dynamic co-word networks to represent sub-topics of social media by dividing the co-word network with community detection algorithms. On the basis of calculating the similarity of sub-topics in different time spans, the evolution of sub-topic is divided into emerging stage, expanding stage and shrinking stage. RE-TE model proposed a sub topic correlation analysis method based on relative entropy, which divided the text stream into text set in continuous time slices, online extracted the latent sub topics in each time slice, based on the online latent Dirichlet allocation model, the sub topics of each time slice are extracted, and the parameters of the model are estimated by Gibbs sampling.

4.3. Experiment 1--Experiments on precision rate and recall rate

In our method, the parameter α plays a crucial role, which controls how many weights the cosine similarity and Jaccard similarity are, respectively. In the extreme case, if we set α to be a tiny value, cosine similarity plays a great role, however, Jaccard similarity is almost omitted, which means that the similarity between social entities is not fully considered. If we set α to be a big value, Jaccard similarity plays a great role, however, cosine similarity is almost omitted, which means that the similarity between feature vectors is not fully emphasized. Therefore, we set α to be 0.5, which not only considers the similarity between feature vectors but also takes into account the similarity between social entities under online social network.

Table 3. Accuracy comparison on both datasets

Dataset	Metrics	CW-TE model	RE-TE model	OTE model
Big data from DBLP	Precision	82.36%	86.92%	92.35%
	Recall	78.45%	81.30%	86.36%
	F-measure	80.36%	84.02%	89.25%
ALS Ice Bucket Challenge from Sina micro-blog	Precision	80.35%	83.91%	89.62%
	Recall	74.45%	79.62%	85.63%
	F-measure	77.29%	81.71%	87.58%

Compared with the CW-TE model and RE-TE model, the experimental results are listed in Table 3. From Table 3, it can be observed that among the baseline methods, our proposed OTE method outperforms CW-TE model and RE-TE model in precision rate, recall rate and F-measure value on two datasets. For example, as to precision rate, our method improves the result by 9.99% and 5.43% compared with CW-TE model and RE-TE model on the dataset big data from DBLP, respectively. For recall rate, the improvements compared to CW-TE model and RE-TE model are 7.91% and 5.06%, respectively. In overall, the improvements of our approach are significant, which justifies the proposed OTE algorithm is effective.

4.4. Experiment 2--Analysis of the intensity evolution of the topics

The two contrast experiment results are used to analyze the intensity evolution of the related topics among the proposed OTE model, CW-TE model and RE-TE model, in which, the parameter of the experiment set up is $\lambda = 0.5$ as shown in Section 4.3.

Figure 2 shows the intensity evolution of topics on ALS Ice Bucket Challenge, regarding the month as time slice and the time span is from May 2014 to May 2015. As can be seen from the evolution of the topic in Figure 2, in most of the cases, OTE are likely to show a better intensity evolution than CW-TE model and RE-TE model in the intensity rising trend from May 2014 to September 2014, or the downward trend from September 2014 to May 2015. Relative to the changes of the other two curves, CW-TE model is a little abnormal, which indirectly shows the disadvantages of CW-TE model, that is, in some cases, if the noise is too much, too many factors would produce the effect amplifying noises. From Figure 2, it can also be seen that the topic on ALS Ice Bucket Challenge is continuous. The intensity change has obvious fluctuation characteristic. The intensity of the topic gradually increased, after a period of time to reach the peak, then the intensity of the topic decreased.This is because that the significant events have sustained high attention, and the frequency of the incident is gradually increasing along with the incident, especially when the event occurs, the report is focused on, which shows the intensity of the topic has risen sharply to the peak. In the latter part of the event, with the degree of concern decreasing, the frequency of the report also fell, which made the intensity of the subject dropped sharply.

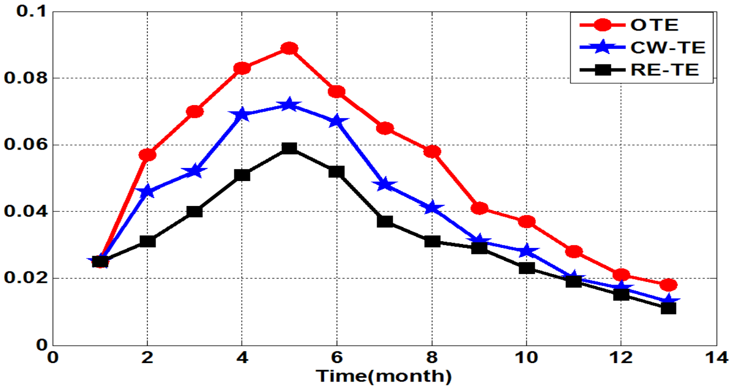

Figure 2. The topic intensity evolution comparison on ALS Ice Bucket Challenge from Micro-blog

Figure 3 shows the topic intensity evolution about the big data information from DBLP from 2005 to 2015. The three curves show a rising trend, which is related to the research arising from data mining and cloud computing after the relational model. From 2005 to 2010, the three curves only have minor changes, which is because that scholars engaging in big data research were still very few. Then since 2011, the topic intensity evolution trends of the three curves are rising rapidly because researches on big data have gradually become hot spots. The trend of the proposed OTE model is relatively large, because of the OTE model simultaneously considers the semantic related feature and topic dynamic evolution on big data, while CW-TE model and RE-TE model are only based on the document and not fully consider the topic dynamic evolution, so that they do not reflect the intensity changes of the topics.

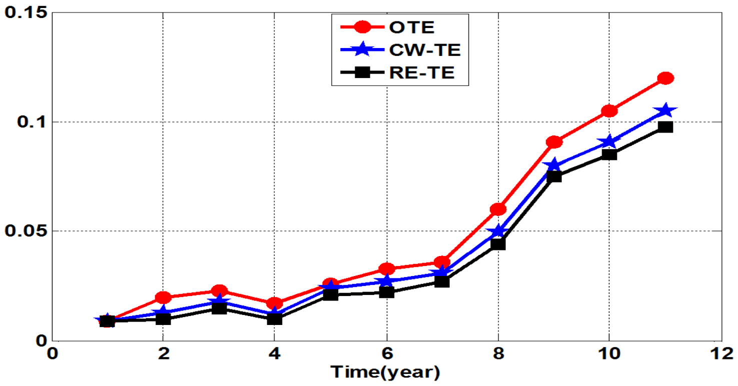

Figure 3. The topic intensity evolution comparison on big data from DBLP

5. Conclusion and future work

Aiming at that the semantic information of nodes under online social network is not considered fully, this paper proposed a new online topic evolution model named OTE method, which is based on the semantic related feature analysis and dynamic evolution under the social network. OTE method comprehensively considers the similarities between key words and between social entities, and K-means algorithm is also adopted to execute the topic clustering. In order to verify the validity of the proposed OTE method, this paper has carried out the experiments on Sina Micro-blog data set and DBLP data set compared with CW-TE model and RE-TE model. Experimental results show that the proposed OTE model has better effect on the topic tracking and dynamic evolution. However, there exist some disadvantages in the proposed OTE method as follows:

(1) Firstly, in this paper, the number of topics is fixed, which cannot detect the disappearance of the existing topic, the generation of new topics and the division of the topic. However, in fact these situations are often happened, which is a very challenging problem in the topic tracking task and needs to be further studied.

(2) Secondly, In our method, the parameter α which controls how many weights the cosine similarity and Jaccard similarity are respectively, is not obtained through the experimental verification but directly given out. Therefore, one of our future researches is how to decide the value of α by the experimental verification.

(3) Thirdly, in future work, we plan to introduce parallel cloud computing platform Hadoop and MapReduce to the proposed OTE method to make the topic intensity evolution analysis more quick and effective.

Acknowledgements

This work was supported by the key project of Anhui Province Colleges and Universities Natural Science Foundation of China (No.KJ2014A250), the project of outstanding young talents supported by Suzhou University of China (No.2014XQNRL009, No.2014XQNRL006, No.2014XQNRL007) and the open project of Intelligent Information Processing Laboratory at Suzhou University of China (No.2014YKF40, No.2014YKF43).

References

1. Saganowski S., Bródka P., Kazienko P. (2012) Influence of the Dynamic Social Network Timeframe Type and Size on the Group Evolution Discovery. *IEEE/ACM Inter. Conf. on Advances in Social Networks Analysis and Mining, IEEE Computer Society*, Istanbul, Turkey, p.p. 678-682.
2. Hu B., Song Z., Ester M. (2012) User Features and Social Networks for Topic Modeling in Online Social Media. *IEEE/ACM Inter. Conf. on Advances in Social Networks Analysis and Mining, IEEE Computer Society*, Istanbul, Turkey, p.p. 202-209.
3. Leskovec J., Lang K.J., Dasgupta A., Mahoney M.W. (2008) Statistical Properties of Community Structure in Large Social and Information Networks. *Proc. on the 17th International Conf. on World Wide Web*, Beijing, China, p.p. 695-704.
4. Tanya Y.B., Jared S. (2006) A Framework for Analysis of Dynamic Social Networks. *Proc. Conf. on the 12th ACM SIGKDD*, Philadelphia, PA, USA, p.p. 20-23.
5. Watts D.J., Strogatz S.H. (1998) Collective Dynamics of 'small-world' networks. *Nature*, 393(6684), p.p.440-442.
6. Brodka P., Musial K., Kazienko P. (2009) A Performance of Centrality Calculation in Social Networks. *Proc. of the 2009 Inter. Conf. on Computational Aspects of Social Networks, IEEE Computer Society*, Washington, DC, USA, p.p. 248-31.
7. Kimura M., Satio K., Motoda H. (2009) Blocking Links to Minimize Contamination Spread in a Social Network. *ACM Transactions on Knowledge Discovery from Data*, 3(2), p.p. 1-23
8. Kwak H., Lee C., Park H., Moon S. (2010) What is Twitter, a Social Network or a News Media? *Proc. Of the 19th Inter. World Wide Web Conference*, Raleigh, NC, USA, p.p. 591-600.
9. Sarkar P., Chakrabarti D., Jordan M.I. (2012) Nonparametric link prediction in dynamic networks. *Proc. of the 29th Inter. Conf. on Machine Learning*, Edinburgh, UK, p.p 245-258.
10. Sun Y.Z., Norick B. (2012) Integrating Meta-path Selection with User-guided Object Clustering in Heterogeneous Information Networks. *Proc. of the 18th ACM SIGKDD Knowledge Discovery and Data Mining*, Beijing, China, p.p 1348-1356.
11. Huang J., Sun H., Han J. (2010) SHRINK: A Structural Clustering Algorithm for Detecting Hierarchical Communities in Networks. *Proc. of the 2010 ACM CIKM Inter. Conf. on Information and Knowledge Management*, Toronto, Canada, p.p. 219-228.

12. Tang J., Sun J., Wang C., Yang Z. (2009) Social Influence Analysis in Large-scale Networks. *Proc. of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Paris, France, p.p 807-815.
13. Chen Z.Q. (2015) Analysis of Topic Evolution in Social Media Based on Co-word Network. *Information Science (In Chinese)*, 33(1), p.p. 120-125.
14. HU Y.L., Bai L., Zhang W.M. (2012) Modeling and Analyzing Topic Evolution. *Acta Automatica Sinica*, 38(10), p.p. 1690-1697.

NCache: A Cache Algorithm of SSD in Storage System

Heng Yang

School of Education Science, Nanyang Normal University, Nanyang 473061, China

Abstract
With the scale of the storage data increasing, the demand for data storage is higher and higher. The performance gap between the central processor and the disk is bigger and bigger so that People have to improve the performance of the disk. This is a big challenge to the storage system. Solid state disk (SSD in short) as a new storage medium, its performance and price are between the central processor and disk. The emergence of solid state disk is an opportunity to the storage system, which can narrow the performance gap between the central processor and disk. It is a hot research field in the current storage area. Based on the characteristics of SSD, we have proposed a new two levels cache algorithm called NCache which is used as a hot data buffer for the system. It can improve the hit rate of the cache and reduce the number of read and write times to improve the performance of the storage system. The buffer of random access memory is divided into the most recently used buffer and write buffer. We used Least Recently Used algorithm (LRU in short), which the latest data is based on to improve the system response time. The data written frequently is stored in the write buffer. The small write data can be aggregated into big write that can shorten the write numbers and extend the life of SSD. Test results show that the design of the two levels cache algorithm can work well. NLRU algorithm has advantages in improving the performance and reducing the number of solid state disk write. And with the use of solid state disk, the system can run faster than the system without SSD, greatly improving the performance of the system.

Key words: TWO LEVELS CACHE ALGORITHM, LRU, SSD, PERFORMANCE

1. Introduction

The explosive growth in social data is a severe challenge to on the traditional storage system. The digital universe Research Report released in the end of 2012 [1] shows that people will produce more than 40ZB data by 2020, which is equivalent to the amount of 5200GB per person on earth. At the same time, the report announces that all the data can be double every