Journal of Engineering and Applied Sciences 12 (10): 2597-2602, 2017
ISSN: 1816-949X

Expert Finding Model Through Author Disambiguation in Bibliographic Data

Jae-Wook Seol, Seok-Hyoung Lee, Seo-Young Jeong, Hye-Jin Lee,
Jeong-Seon Yoon and Kwang-Young Kim
Korea Institute of Science and Technology Information (KISTI), Seoul, Korea

Abstract: In the modern society, unexpected events such as diseases and disasters are advertent. In specific specialized sector, finding of expert is important for resolving social issues. This study proposes that expert finding model for each sector through quantification of expertise of researchers. First, the issue of ambiguity of author in academic data shall be resolved. To measure precisely the expertise of an author we conduct author identification in which the author name written in different forms is identified as an actual personnel. Second, based on the accumulated importance of author keyword and reference network, we apply the modified Hyperlink-Induced Topic Search (HITS) algorithm to extract out expert candidates. To verify the validity of this proposal, expert finding shall be conducted on 92,100 cases of academic data incurred in Korea. We evaluate our expert finding model based on human relevance judgments about several queries. The outcome of experimenting author importance resolution is F1 measure 94.79% and the expert finding model applied with our modified HITS algorithm shows the mean average precision of 75%.

Key words: Expert finding, HITS algorithm, author disambiguation, support vector machine, identification

INTRODUCTION

In the modern society recently, unexpected events such as diseases including the MERS-CoV and Ebola virus as well as disasters including the capsized Sewol ferry are frequently occurring. Incidents not prepared for result in massive damage in property and lives. To resolve these events we need expert advice. Therefore, expert finding in specific expertise sector is crucial in resolving social issues. Recently, the expert finding researches are being actively conducted in social network services (Zhang *et al*., 2007) and online knowledge communities (Wang *et al*., 2013). However, previous researches have been limited to specific areas for implementation of expert finding. The purpose of expert finding is to find the expert in the relevant area in a specific situation. In this aspect this study using these data set throughout sectors can find expert over diverse sectors.

This study proposes expert finding model based on abundant bibliographic data through quantification of expertise of researchers in diverse sectors. We extracted expert candidates by applying the modified HITS algorithm based on the accumulated importance of authority keyword and reference network. To verify the validity of expert finding model, expert finding was conducted on 287,352 researchers and 92,100 cases of academic data created in Korea. However, the only problem with our test dataset is that there is an author ambiguity. To measure the expertise of author precisely we have to identify the author name in different forms as actual personnel.

Author identification, a way of resolving author ambiguity is by classifying authors with the same name into actual personnel. Author identification groups the same author into one cluster. Author disambiguation task, also, known as authorship verification, authorship analysis and author clustering is being actively researched on. In particular, PAN has been intensively researching from 2011, until today through author identification task. In the previous researches, the properties used for author identification were commonly the affiliated institutions, thesis title, keyword, author and published year. This is because the information available from academic data is limited. Likewise, insufficient property information leads to difficulty in author identification and limitation to improvement of identification precision. To overcome this limited information we propose the method of increasing author identification performance by expanding co-author network.

Literature review: PAN has intensively researched from 2011, until today in the author identification area. PAN 2015 task is the most similar to our author identification task. In this task, the classification approach and feature used for author identification are the key elements. All

Corresponding Author: Jae-Wook Seol, Korea Institute of Science and Technology Information (KISTI), Seoul, Korea

participants used supervised learning method. Most participants used the well known machine learning called Support Vector Machine (SVM) (Giles *et al*., 2005; Li *et al*., 2007) and also, used the decision trees and random forests (Palomino-Garibay *et al*., 2015). By using the non-instructive learning method not used in the PAN task, author identification was attempted. As the unsupervised learning technique, the author identification was conducted by using HAC (Hierarchical Agglomerative Clustering) (Yang *et al*., 2011; Aswani *et al*., 2006; Torvik *et al*., 2005; Ikeda *et al*., 2009) K-means and DBSCAN (Huang *et al*., 2006) by extracting the similarity between author sets.

By using various properties for author identification, the performance can be enhanced. The properties used for author identification in previous researches can be largely divided into the intrinsic features and extrinsic features. Intrinsic features are the properties written in the data, for example, the e-mail address, co-author name, thesis title, abstract and keyword. Extrinsic features are the properties of data information that can be gained from web or dictionary search (Kang *et al*., 2009; Kanani and McCallum, 2007) for example, the publication list of author (Yang *et al*., 2006), curricula vitae, mesh term (Treeratpituk and Giles, 2009), document similarity and word similarity (Yang *et al*., 2011). To use various extrinsic features, access to information is limited and it takes a lot of time for web search on data. In this study, the items in the bibliography of thesis shall be used as features and the author name identification issue will be efficiently handled by expanding the given bibliography information.

Expert finding based on social network or relation network calculates the authority value of user by analyzing the established link. From the calculated authority value, the user top-k by area and sector is extracted. Li *et al*. (2007) and Zhang *et al*. (2007) established network through the co-authorship of authors from the researcher related data gathered from the web. Based on the established network, the relevancy of people on topic q was calculated. Bouguessa *et al*. (2008) established the network of questioners and answerers from the data gathered from Yahoo Answers. They address to model the authority scores of users as a mixture of gamma distributions. The number of components in the mixture is estimated by the Bayesian Information Criterion while the parameters of each component are estimated using the Expectation-Maximization algorithm. Our expert finding method uses link analysis by establishing citation network from the citation in theses from academic data and accumulated importance of author keywords to calculate the academic expertise of authors.

MATERIALS AND METHODS

Disambiguating author and modeling expert finding: In this study, first we deal with the author ambiguity issue in academic data. To measure precisely the expertise of author we conducted author identification to identify the author names written in different forms into an actual personnel. Second, based on the accumulated importance of author keyword and reference network we extract out expert candidates by applying the modified HITS algorithm.

Author disambiguation using expanding co-author network: To judge if two arbitrary people with the same name are the same person, it is important to find out features as clues. The features used for author identification previously include title, keyword, co-author and affiliation. Among these features, co-author is the most intuitive and effective feature for author identification. Yang *et al*. (2008) said that co-author provides the most useful information for author identification. For author identification we expand the co-author network by using co-author information which is the most intuitive and accurate.

Table 1 shows an example of bibliographic data corresponding to "Tae-Sung Kim." If we group all the authors with name "Tae-Sung Kim," the same name group corresponding to "Tae-Sung Kim" will be constructed as A1-A4 and so on.

Figure 1 shows how to expand the co-author network. The co-authors of author A1 are C1-C5. The co-authors of author A2 are C4-C7. The common

Table 1: Example of test data

Data	Text
Data 1	
Author	Tae-Sung Kim (A1)
Co-author	Hee-Sun Kim; Seung-Hun Jin;
Affiliation	Department of Computer Science, Hanyang University
Mail	taesung@hy.ac.kr
Keywords	lactamase inhibitor; *Pseudomonas* sp.;
Data 2	
Author	Tae-Sung Kim (A2)
Co-author	Se-Jin Oh; Han-Kyu Choi; Seog-Tae Han;
Affiliation	Dept. of Industrial Engineering, Seoul University
Mail	tskim@snu.ac.kr
Keywords	quality assurance;
Data 3	
Author	Tae-Sung Kim (A3)
Co-author	Seung-Hun Jin; Ho-Won Kim;
Affiliation	Department of Computer Science, Hanyang University
Mail	taesung@hy.ac.kr
Keywords	*Pseudomonas* sp.;
Data 4	
Author	Tae-Sung Kim (A4)
Co-author	Se-Jin Oh; Seog-Tae Han;
Affiliation	Dept. of Industrial Engineering, Seoul University
Mail	tskim@snu.ac.kr
Keywords	NONE

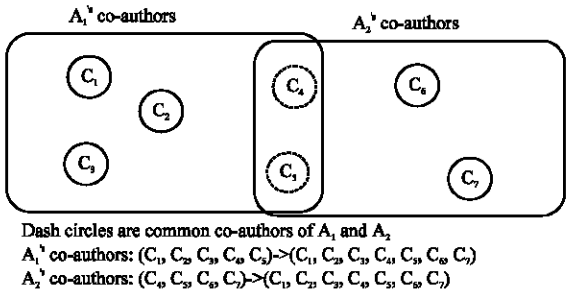

Fig. 1: Expanding co-author network

co-authors of authors A1 and A2 are C4 and C5. If common co-author of A1 and A2 using the same name exists then by adding the co-author of the same name one does not have as the co-author of oneself, the co-author network can be expanded. From this process, author A1 can add co-authors C6 and C7 to the existing co-author group. Likewise, author A2 adds co-authors C1-C3 to the existing co-author group.

However, there is a possibility that the common co-authors with the same name are people with the same name. For example in Fig. 1, C4 is a co-author of A1 and A2 but in the real world, he/she may be a different person with the same name. In this case, a wrong network is formed resulting in error in author identification. The conditions for common co-author for preventing this are as follows. The conditions are divided into when there are two or more common co-authors and when there is one common co-author. First in case, common co-authors are two or more as there is no possibility of error explained above they can be applied as common co-authors. This is because there has been no case where two or more common co-authors were different people after observing the data. Second, in case there is one common co-author, apply as a common co-author if the affiliation matches; otherwise, do not apply as a common co-author.

Table 2 shows an example of author identification using the co-author network expansion on four people (Auth1-Auth4) with the same name called "Doo-Su Lee". While, we can intuitively know that Auth1 and Auth4 are not the same person, when we identify authors by using the co-author network expansion, we can identify them as the same author. This can be predicted as the co-authors of Auth3, "T.S. Jeong" and "T.J. Jeong" (co-authors of Auth1 and Auth3) and "S.G. Lee" and "J.H. Jo" (co-authors of Auth4 and Auth3) exist. However, if we simply use co-author we couldn't be able to identify them as the same author as the common co-authors of Auth1 and Auth4 do not exist.

Table 2: Example of author identification using co-author network expansion

ID	Auth1	Auth2	Auth3	Auth4
Name	Doo-Su Lee	Doo-Su Lee	Doo-Su Lee	Doo-Su Lee
Affiliation	A	B	A	B
Publication Year	1997	2204	-	2001
Co-authors	Y.T. Kim	Y.T. Kim		
	T.S. Jeong		T.S. Jeong	
	T.J. Jeong		T.J. Jeong	
			S.G. Lee	S.G. Lee
			J.H. Jo	J.H. Jo
Cluster (answer)	C1	C1	C1	C1

However, there are limitations to identifying authors in this method alone. For example, if the two individual authors of the same name do = not have = a common co-author, author identification is impossible. For this reason, we need to expand the author identification coverage by using other bibliographical information other than the co-author information.

The clues used for identifying authors of the same name in academic data into personnel in the real world are called the features. In this study, we shall use affiliation, major, e-mail and keywords as the features. Affiliation is the institution affiliated by the author at the time of writing the thesis. Affiliation may include school names, company names and research institutions. This can be good information for determining same persons if there are not many people of the same names in the institutions. Major is the university major of the author. Major can include the department or faculty name. This can be good information for determining same persons if the major area of author does not change. The e-mail is e-mail address stated by the author in the thesis. This is unique information symbolically representing the author. Most people use the same ID on different domains; hence, author identification by extracting ID from e-mail address can be possible. Lastly, keyword is the keyword on research stated in the thesis. Based on the assumption that most researchers research on similar subjects to the major in close period this can be useful information in author identification.

SVM classification is known as the most general and efficient classification in classifying the data for resolving problems in various sectors (Yang *et al*., 2006; Treeratpituk and Giles, 2009). We will identify authors in test set through learning model after learning by using the LIBSVM (Yang *et al*., 2011) of the academic data given from the co-author network proposed in Clause 3.2 and four features proposed in Clause 3.3.

Table 3 shows an example of learning sample on an arbitrary pair of two authors from the same name group of "Tae-Sung Kim". Five number fields composed in blank units in the training sample refer to "same person of

Table 3: SVM learning sample extracted from Table 1

Author entity pair	Training sample					
(A1, A2)	0	0:0	1:0	2:0	3:0	4:0
(A1, A3)	1	0:3	1:1	2:1	3:1	4:1
(A1, A4)	0	0:0	1:0	2:0	3:0	4:0

pair of two authors (answer)", "number of common co-authors", "compliance of institution (match: 1, non-match: 0)", "compliance of major (match: 1, non-match: 0", "e-mail compliance (match: 1, non-match: 0" and "availability of common keyword (present: 1, absent: 0)", respectively. For example, from the pair of authors (A1 and A3), number 1 in the first field represents that A1 and A3 are the same person by SVM class. Number 3 in the second field (0:3) represents the number of common co-authors ("Se-Jin Oh", "Han-Kyu Choi" and "Seong-Tae Han"). The number 1 in third field (1:1) represents that the institution ("Hanyang University") matches. The number 2 in fourth field (2:1) represents that the major ("Computer Science") matches. The number 2 in fifth field (3:1) represents that the e-mail ("taesung@hy.ac.kr") matches. And the number 1 in the sixth field (4:1) represents that a common keyword ("*Pseudomonas* sp.") exists.

Expert finding model based on modified hits algorithm: To extract authoritative expert, the accumulated importance of author keyword and reference network were applied to HITS algorithm (Kleinberg, 1999). HITS algorithm is a link analysis algorithm based on the link indicating another document from a document. By using the network between documents in the research result on query language, the most "authoritative" document is found for appropriate search result.

In this study to apply the modified HITS algorithm, the query language, document and link information were applied as sector name, researcher thesis and reference network, respectively. The below pseudo code is our modified HITS algorithm. The weight of each researcher is set as follows:

$$\text{Auth}^{t+1}(v_i) = \sum_{j:e_{ij}\in E} w_{ji} \times \text{Hub}^{t}(v_j) \quad (1)$$

$$\text{Hub}^{t+1}(v_i) = \sum_{j:i_{ij}\in E} w_{ij} \times \text{Auth}^{t}(v_j) \quad (2)$$

Modified HITS algorithm:

```
A: = set of authors
w: = weight of authors
q: = query
k: = Iteration
function ModifiedHITS (A)
  for each author a in A do
   a.auth = a.avgRef×a.numKwd
   a.hub = 1
  for step from 1 to k do
   norm = 0
   from each author a in A do
    a.auth = 0
    for each author b in a.InCitationAuthors do
     a.auth+ = b.hub×wba
    norm+ = square (a.auth)
   norm = sqrt (norm)
   for each author a in A do
    a.auth = a.auth/norm
   norm = 0
   for each author a in A do
    a.hub = 0
    for each author c in a.OutCitationAuthors do
     a.hub+ = c.auth×wab
    norm+ = square (a.hub)
   norm = sqrt (norm)
   for each author a in A do
    a.hub = a.hub/norm
```

The authority and hub score of each node iteratively updates the scores until convergence according to the modified HITS algorithm. The method selects authors whose authority score are 0.1 or more as an influential supporter. This is because a threshold is set to 0.1 empirically. The initial authority score of an author is set as follows:

$$\text{Auth}^{0}(v_i) = \text{avgRef}(v_i).\ \text{numKwd}(v_i, k) \quad (3)$$

$$\text{Hub}^{0}(v_i) \quad (4)$$

where, avgRef(v_i) represents the number of average citations used by other these from the thesis written by author v_i. The avgRef (v_i) is the level of influence and reliance of author v_i on other researchers with higher value representing better quality of thesis. The numKwd (v_i, k) is the number of documents including the keyword k among the theses written by author v_i. The numKwd (v_i, k) represents the research areas of interest of the author. The more keyword k, the higher interest in research on keyword k. Also, the initial score of the hub score are set as follows: $\text{Hub}_0(v_i) = 1$.

RESULTS AND DISCUSSION

Experiments

Data set: To test the author disambiguation through expansion of co-author network as proposed, the 92,100 datasets published in Korea were experimented. Table 4 shows the bibliographic data set for testing author identification. The total number of authors in 92,100 theses is 287,352. Of all authors, the number of authors in the same name group with same names is 53,526. The actual number of authors excluding any overlaps is 103,559.

Table 4: Bibliographic data set

No. of papers	No. of author entities	No. of same name author groups	No. of real authors
92,100	287,352	53,526	103,559

Table 5: Results from feature contributions for the author disambiguation

Methods	Precision (%)	Recall (%)	F1-measure (%)
(A) co-author (baseline)	56.43	50.00	53.00
(B) A+keyword	98.65	78.89	87.72
(C) B+e-mail	98.66	79.03	87.76
(D) C+major	97.19	87.68	92.19
(E) D+affiliation	96.06	89.12	92.45
(F) co-author network (Proposed method)	94.78	94.80	94.79

Assessments: To measure the performance of author identification we need to judge the conformity of people of the same name in the same name group. From the same name group, the number of arbitrary pairs of two authors is 2,124,400. The 2/3 and 1/3 of the data are composed of training set and test set, respectively and were evaluated by 3-fold cross validation.

The correctness of system result was determined by the SVM result of arbitrary pair of two authors in the same name group and the conformity of the answer set established previously.

We evaluated systems using recall, precision and the F1-measure (Eq. 1-3). These metrics rely on True Positives (TP), False Positives (FP) and False Negatives (FN) which are defined as appropriate in order to provide exact and inexact evaluation of the tasks:

- Precision (P) = TP/(TP+FP)
- Recall (R) = TP/(TP+FN)
- F1-measure (F) = 2×P×R/(P+R)

As there is no correct answer set for expertise, the performance of expert finding system is hard to measure. To verify the performance of expert finding system we utilized the web portal (Naver) and major personnel search system in each sector (Joins). In case, the personnel search result falls on either of the following three cases we determined as expert: first, if a news article related with relevant field comes out when searched the personnel in search result on web portal, second, if this person is found to be a popular person in web portal and third, if the search result of the relevant person comes out from searching on Joins.

Experimental results: Table 5 shows the result from combination of features for author identification. Each method is how features are combined to confirm their contributions for the task. Our baseline used only co-author feature and we denote it as (A) baseline. Then, the results are listed by adding the features keyword (B), e-mail (C), major (D) and affiliation (E), respectively.

Table 6: Results from proposed method for the author disambiguation

Methods	Precision (%)	Recall (%)	F1-measure (%)
1-fold	94.00	94.12	94.06
2-fold	95.40	95.43	95.41
3-fold	94.93	94.86	94.89
Averages	94.78	94.80	94.79

Table 7: Expert finding system results by field

Fields	P@10	P@15	P@20	MAP
Machine learning	0.60	0.53	0.50	0.71
Radiation therapy	0.70	0.66	0.70	0.73
Finite element analysis	0.80	0.80	0.80	0.89
Acupuncture	0.60	0.53	0.50	0.71
Kimchi	0.80	0.86	0.70	0.71
Average	0.70	0.67	0.64	0.75

Our proposed method (F) is by combining the four features except the co-author feature (A) with co-author network.

Table 6 shows the result of implementing author disambiguation by using the proposed method of dividing the 1/3 of the overall data into test sets. The proposed method was effective in identifying authors with 94.79% average F1 measure.

The proposed author identification through expansion of co-author network showed the most effectiveness than any other combinations of features. This is because an additional clue for identifying authors was found by expanding the given information from academic data. This has significance in that it expands the limited information to maximize the identification performance.

Table 7 shows the expert finding results in five sectors. We represented the search results on each keyword into precision at k (P@k) and Mean Average Precision (MAP). From the expert finding experiment result by sector, the average of P@10 was 70% with 75% performance of MAP. We attempted to resolve the author ambiguity issue before conducting the expert finding system. If we did not resolve the author ambiguity issue, the performance on authors will be dispersed. This issue will interfere proper reflection of author influence when measuring the author expertise.

CONCLUSION

In this study, the authors of same name from academic data were identified and expert in each sector was extracted based on the identification result. For author identification we have established the co-author network and based on this as the SVM feature, 94.7% (F) of performance was gained from effective author identification. To extract expert in each sector we calculated the expertise by using the number of citations in theses of researchers and number of keywords and applied these as the weighted value of HITS algorithm. From the expert finding experiment result on five sectors, MAP showed 75.0% performance.

SUGGESTIONS

For future work, we will reflect the time graph and co-author network for expert finding. In this study, expert finding did not take into consideration the recency and researcher network in each sector. To reflect the latest issue on each sector we will reflect the time graph and for recommendation of expert with the similar tendency, the co-author network will be utilized. Also, we will add reports and patent data along with the theses data used for the experiment to implement the author identification and expert finding system.

REFERENCES

Aswani, N., K. Bontcheva and H. Cunningham, 2006. Mining information for instance unification. Proceedings of the 5th International Conference on Semantic Web (ISWC 2006), November 5-9, 2006, Springer, Athens, Georgia, USA., pp: 329-342.

Bouguessa, M., B. Dumoulin and S. Wang, 2008. Identifying authoritative actors in question-answering forums: The case of yahoo! answers. Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 24-27, 2008, ACM, New York, USA. isBN:978-1-60558-193-4, pp: 866-874.

Giles, C.L., H. Zha and H. Han, 2005. Name disambiguation in author citations using a k-way spectral clustering method. Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'05), June 7-11, 2005, IEEE, Denver, Colorado, USA., ISBN: 1-58113-876-8, pp: 334-343.

Huang, J., S. Ertekin and C.L. Giles, 2006. Efficient name disambiguation for large-scale databases. Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases Vol. 6, September 18-22, 2006, Springer, Berlin, Germany, pp: 536-544.

Ikeda, M., S. Ono, I. Sato, M. Yoshida and H. Nakagawa, 2009. Person name disambiguation on the web by two-stage clustering. Proceedings of the 2nd Workshop on Web People Search Evaluation (WePS 2009), April 20-24, 2009, World Wide Web, Madrid, Spain, pp: 1-6.

Kanani, P. and A. McCallum, 2007. Efficient strategies for improving partitioning based author coreference by incorporating Web pages as graph nodes. Proceedings of the AAAI 2007 Workshop on Information Integration on the Web, July 23, 2007, AAAI Press, California, USA., pp: 38-43.

Kang, I.S., S.H. Na, S. Lee, H. Jung and P. Kim *et al*., 2009. On co-authorship for author disambiguation. Inf. Process. Manage., 45: 84-97.

Kleinberg, J.M., 1999. Authoritative sources in a hyperlinked environment. J. ACM, 46: 604-632.

Li, J., J. Tang, J. Zhang, Q. Luo and Y. Liu *et al*., 2007. Eos: Expertise oriented search using social networks. Proceedings of the 16th International Conference on World Wide Web, May 8-12, 2007, ACM, New York, USA. isBN:978-1-59593-654-7, pp: 1271-1272.

Palomino-Garibay, A., A.T. Camacho-Gonzalez, R.A. Fierro-Villaneda and I. Hernandez-Farias *et al*., 2015. A random forest approach for authorship profiling. Masters Thesis, University of Namibia, Windhoek, Namibia.

Torvik, V.I., M. Weeber, D.R. Swanson and N.R. Smalheiser, 2005. A probabilistic similarity metric for Medline records: A model for author name disambiguation. J. Assoc. Inf. Sci. Technol., 56: 140-158.

Treeratpituk, P. and C.L. Giles, 2009. Disambiguating authors in academic publications using random forests. Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries, June 15-19, 2009, ACM, New York, USA. isBN:978-1-60558-322-8, pp: 39-48.

Wang, G.A., J. Jiao, A.S. Abrahams, W. Fan and Z. Zhang, 2013. Expert rank: A topic-aware expert finding algorithm for online knowledge communities. Decis. Support Syst., 54: 1442-1451.

Yang, K.H., H.T. Peng, J.Y. Jiang, H.M. Lee and J.M. Ho, 2008. Author name disambiguation for citations using topic and web correlation. Proceedings of the 12th European Conference on Research and Advanced Technology for Digital Libraries, (ECDL 2008), September 14-19, 2008, Springer, Aarhus, Denmark, pp: 185-196.

Yang, K.H., J.Y. Jiang, H.M. Lee and J.M. Ho, 2006. Extracting citation relationships from web documents for author disambiguation. Institute of Information Science, Academia Sinica, Taipei, Taiwan.

Yang, X., P. Jin and W. Xiang, 2011. Exploring word similarity to improve chinese personal name disambiguation. Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT) Vol. 3, August 22-27, 2011, IEEE, Lyon, France, ISBN:978-1-4577-1373-6, pp: 197-200.

Zhang, J., J. Tang and J. Li, 2007. Expert finding in a social network. Proceedings of the 12th International Conference on Database Systems for Advanced Applications (DASFAA 2007), April 9-12, 2007, Springer, Bangkok, Thailand, pp:1066-1069.