Author Name Disambiguation for Ranking and Clustering PubMed Data Using NetClus

Arvin Varadharajalu[1], Wei Liu[1,⋆], and Wilson Wong[2]

[1] School of Computer Science and Software Engineering
The University of Western Australia, Australia
[2] School of Computer Science and Information Technology,
RMIT University, Australia
wei@csse.uwa.edu.au

Abstract. The ranking and clustering of publication databases are often used to discover useful information about research areas. NetClus is an iterative algorithm for clustering heterogenous star-schema information network that incorporates the ranking information of individual data types. The algorithm has been evaluated using the DBLP database. In this paper, we apply NetClus on PubMed, a free database of articles on life sciences and biomedical topics to discover key aspects of cancer research. The absence of unique identifiers for authors in PubMed introduces additional challenges. To address this, we introduce an improved author disambiguation technique using affiliation string normalisation based on vector space model together with co-author networks. Our technique for disambiguating authors, which offers a higher accuracy than existing techniques, significantly improves NetClus clustering results.

Keywords: Author Disambiguation, Clustering, NetClus, Heterogeneous Information Network.

1 Introduction

Governments, businesses, pharmaceutical companies and individual researchers frequently search publication databases to find the leading experts and their research groups [3]. PubMed is a free database accessing primarily the MEDLINE database of references and abstracts on life sciences and biomedical topics. Clustering query results of such databases will uncover invaluable information that would otherwise not available. For example, publications about ongoing research on cancer can be retrieved, and cluster analysis can be applied to answer questions such as new research findings, causes of cancer, common types of cancer and their treatment, as well as who the leading researchers and institutions are. However, PubMed's data collection is driven by crawling. In other words, authors and institutions are not assigned with unique identifiers as in the case of curated databases such as DBLP. Moreover, the increase of biomedical research citations in PubMed in the recent years makes the task of manually converting the literature into structured data extremely difficult [3].

⋆ Corresponding author.

D. Wang and M. Reynolds (Eds.): AI 2011, LNAI 7106, pp. 152–161, 2011.

Publication data is best represented as a heterogeneous information network, which is a data model representing relations among multiple types of objects. RankClus [5] and its improvement NetClus [6] are among the more popular algorithms that integrate ranking and clustering for bi-typed and multi-typed heterogenous information networks, respectively. Both algorithms recognise the important fact that the ranking and the clustering mutually enhance each other. Ranking objects without clustering will lead to incomplete results, e.g., ranking artificial intelligence and database conferences together may not make much sense. Similarly, clustering a vast number of data objects in one huge cluster without differentiation is not informative either. However, to apply NetClus on PubMed, the ambiguity in author and institution names has to be resolved.

In this paper, we propose a technique to disambiguate authors by normalising affiliation string using a vector space model and then combining with co-author networks to further improve the system performance. The problem of affiliation string normalisation, especially on the PubMed dataset, has not been adequately addressed. Author disambiguation relying only on the affiliation strings suffers from the problems of variation in affiliation information, non-standard representation of affiliations, multiple affiliations for the same author and so on. We apply NetClus on PubMed publication records after disambiguation using our technique to extract useful information to understand the research trend, the leading journals, and organisations in the field of cancer research. To achieve this, we (1) developed a software module to collect PubMed records; (2) designed and developed the technique for author disambiguation; (3) reviewed existing clustering algorithms to identify the most suitable for this task; and (4) adapted the NetClus algorithm to work with PubMed data.

The paper is organised as follows. Section 2 examines the field of author disambiguation to identify the special focus of our disambiguation techniques. Section 3 outlines our disambiguation technique. In section 4, our algorithm is evaluated and clustering results are presented. A comparison of NetClus clustering results with and without author disambiguation is also presented. The paper concludes in Section 5 with an outlook to future work.

2 Related Work

2.1 The Challenges of Author Name Disambiguation on PubMed

The author name *"Wei Zheng"* is very common in PubMed articles. More than 1,700 publications were retrieved when *"Zheng W"* [Author] was searched on 10th of May, 2011. According to the PubMed search interface, all these articles were published by a single author but in actual fact, all these articles are published by authors with same name. This ambiguity problem leads to poor clustering results and incorrect co-author networks. Unlike DBLP, there is no unique ID that can identify an author or an institution. More sophiscated interfaces for PubMed such as GoPubMed[1] is able to offer multi-faceted search

[1] `http://www.gopubmed.org/`

experience such retrieving the affiliations and basic statistics about an individual author. These interfaces, however, are far from accurate and reliable. Just like PubMed, GoPubMed suffers the same problem of author name ambiguity, which messes up all the otherwise useful statistics.

Affiliation is probably the simplest starting point for author disambiguation. However, no standard format is enforced when affiliations are attached to an article in PubMed. Therefore the huge variation in affiliation strings makes it difficult to classify authors. *"Baylor College of Medicine"*, in the example below, is taking on a variety of names. `s1` has author name in it. `s2` and `s3` have email address but `s3` is missing the department. `s4` contains three different departments.

```
s1: Texas Children's Cancer Center, Michael E DeBakey Department of
    Surgery, Baylor College of Medicine, Houston, Texas 77030, USA.
s2: Texas Children's Cancer Center/Baylor College of Medicine,
    Houston, Texas 77030, USA. tmhorton@txccc.org
s3: Baylor College of Medicine, Houston, Texas 77030-2399, USA.
    dmetry@bcm.tmc.edu
s4: Texas Children's Cancer Center, Department of Pediatrics, Dan L.
    Duncan Cancer Center, Baylor College of Medicine, Houston, TX, USA.
```

To complicate matters, one author may have multiple affiliations at the same time, or at different times when they move between institutions, as illustrated by the author *"John M Maris"* below:

```
s1: Texas Tech University Health Sciences Center, Lubbock,
    Texas. min.kang@ttuhsc.edu.
s2: St. Jude Children's Research Hospital, Memphis, Tennessee.
```

2.2 Related Work on Disambiguation of PubMed Authors

Many techniques have been developed for disambiguating author names from PubMed articles. Torvik et al. [7], for instance, used a combination of features such as title words, journals, medical headings to calculate a probability to determine whether two articles have the same authors. This technique is based mainly on the assumption that authors publishing papers in similar research areas are more likely to be the same. However, the granularity and subjectivity of associating a publication with a research area may break this assumption.

Yu et al. [9] used the connections between organisations to articles and authors to disambiguate authors using affiliation strings. They disambiguate authors by extracting organisations and related entities from affiliation strings. There are, however, several problems with this technique. First, they assumed that an affiliation string follows this format: (address component, address component, country, email). This is never always true as shown in the example below:

```
s1: Integrated microRNA and mRNA expression profiling in a rat colon
    carcinogenesis model: Effect of a chemo-protective diet.[21406606]
    [Affiliation : Texas AM University]
s2: An algorithm to detect a center of pupil for extraction of point
    of gaze. [Dept. of Biomed. Eng., Inje Univ., Kimhae, South Korea.]
```

Second, dictionaries are required to differentiate geographical information from the names of organisations in the address component. Third, it requires that the organisation names contain only English words. The fact is, however, almost 10% of the affiliations contain words from the authors' native languages [3].

Jonnalagadda et al. [3] have taken these drawbacks into consideration and developed a technique called NEMO. They assumed that the information available in affiliation string is sufficient enough to disambiguate authors. In reality, many publications in PubMed use the affiliations of the first authors only. The affiliation strings therefore cannot always be used to disambiguate all authors.

3 A Multi-evidence Author Disambiguation System

There are two parts in our proposed technique as shown in Figure 1, namely, the extraction of different components from affiliation strings, and the 3-phase multi-evidence disambiguation process.

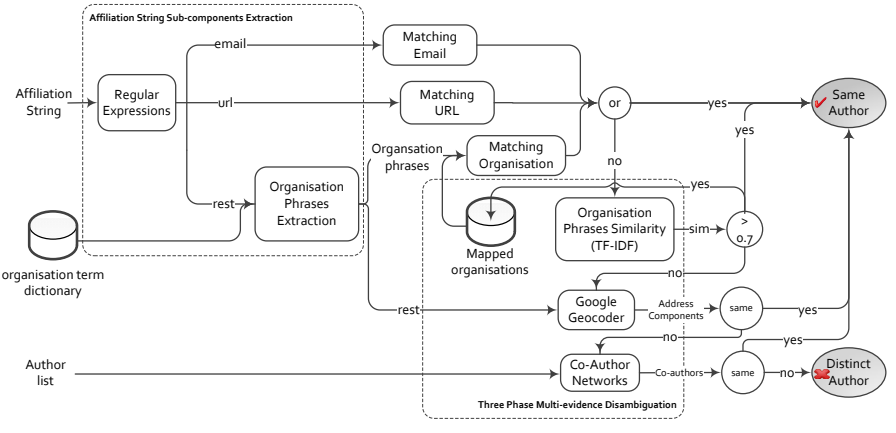


Fig. 1. Conceptual Overview of the Name Disambiguation System

In the first part, the affiliation strings of an ambiguous name (i.e. a name that can potentially refer to multiple individuals) are broken down into different components, namely, organisation names, organisation addresses, and email or homepage addresses, using a combination of regular expressions and other tools. The email address and URLs are first extracted using the following expressions:

```
/[\textbackslash.\_a-zA-Z0-9\-]+@[\textbackslash.\_a-z A-Z0-9-]+/i

((https?|ftp|gopher|telnet|file|notes|ms-help):((//)|(\textbackslash
\textbackslash\textbackslash\textbackslash))+[\textbackslash w
\textbackslash d:\#@\%/;\$()\~\_?\textbackslash+-=\textbackslash
\textbackslash\textbackslash.\&]*)
```

The remaining strings are then fed into a module which extracts the names of organisations using the following resources and tools:

1. Abbreviations and acronyms are disambiguated using the service provided by websites such as `Abbreviations.com`
2. Typographical mistakes are corrected using edit-distance [2] as in [8].
3. A table of organisation names as in [3].

The detected email addresses and URLs as well as organisations names are then removed from the affiliation strings. Finally, using Google's Geocoder[2], the address components, namely, countries, states, cities and street names, are extracted from the words that remain in the affiliation strings.

In the second part, the extracted components are used to determine if an author with multiple affiliation strings actually refer to one or multiple individuals. Initially, if the email addresses or URLs from two affiliation strings match, the two corresponding authors are considered to be the same. Otherwise, the technique proceeds further to the next two phases.

3.1 Disambiguation Using Organisation Names and Addresses

In this second phase, the organisation name components of the affiliation strings are compared using a vector space model [1]. Every term in the organisation name is represented as a dimension, and TF-IDF [4] is used to compute the weights of the individual terms. In our technique, the threshold value is set to 0.7. In other words, to be considered as identical organisations, the cosine similarity needs to be greater than 0.7. In the following example, the one author name has four different affiliation strings:

```
s1: Department of Molecular and Cellular Biology and Dan L. Duncan
    Cancer Center, Baylor College of Medicine
s2: Lester and Sue Smith Breast Center, Baylor College of Medicine
s3: Paris Breast Center, L'Institut Du Sein
s4: Clinical Research Division, Fred Hutchinson Cancer Research Center
```

All stop words are removed to improve the results. The term frequency, document frequency, inverse document frequency and term weights are then calculated. The word *"baylor"*, for instance, appears only once in the first organisation name. Therefore the term weight is assigned only to the first string. The word *"center"* appears in all the strings therefore its weight is reduced to zero using IDF. Similarly, the term weight for each and every term is calculated by multiplying the number of strings containing the term with the inverse document frequency. Next the dot products of all possible pairings of strings are calculated. If the cosine value is 1, then the strings within a pair are identical. In this example, the cosine values between $s1$, and $s2$, $s3$ and $s4$ are 0.6345, 0 and 0.1502, respectively. In other words, the second organisation name $s2$ is the most similar to $s1$. However, in this example, the cosine value is still less than 0.7. The chances

[2] `http://code.google.com/apis/maps/documentation/geocoding/`

of the authors of both strings being the same are very high. This problem is resolved by considering the address components in the affiliation strings. If three out of the four components (i.e. country, state, city, street name) match, the two organisation names and hence, the authors are considered to be the same. If this step fails, we proceed to the next step involving co-author networks.

3.2 Disambiguation Using Co-author Network

In this phase, the technique deals with a single author that has completely different email addresses, URLs, organisation names and organisation addresses. The author *"John M Maris"* mentioned at the end of Section 2.1 is an example. In cases like this, the technique will proceed to examining the co-author network to identify whether the multiple *"John M Maris"* are actually referring to the same or different individuals given these affiliation strings.

```
s1: Min H Kang, C Patrick Reynolds, Peter J Houghton, Denise Alexander,
    Christopher L Morton, EAnders Kolb, Richard Gorlick, Stephen T Keir,
    Hernan Carol, Richard Lock, John M Maris, Amy Wozniak, Malcolm A Smith
s2: Christopher L Morton, John M Maris, Stephen T Keir, Richard Gorlick,
    E Anders Kolb, Catherine A Billups, Jianrong Wu, Malcolm A Smith,
    Peter J Houghton
```

Using the co-author path from Microsoft Academic Search, the technique is able to determine that both *"John M Maris"* refer to the same author in the two publications above as *"John M Maris"*, *"Christopher L Morton"*, *"Malcom A Smith"* and *"Peter J Houghton"* have co-authored paper previously.

4 Evaluation of the Disambiguation Technique

For this evaluation, we queried PubMed using the word *"cancer"*. 12,707 results were retrieved between the year 2011 and 2010. From a total of 12,700 articles, our system has identified 80,042 paper-author combinations, and 71,810 unique authors. Table 1(a) lists the top 12 ambiguous names and the actual number of authors they represent. Out of the unique authors, 6,152 authors had more than two publications. Therefore, 65,658 authors have published only one paper in this dataset. From the 6,152 authors with multiple publications, 5,203 were identified using co-author network. The chance of these authors being different is therefore highly unlikely. Therefore, we only need to verify the correctness of the remaining 949 authors. These 949 authors' article titles and abstract were extracted. The entire code of this work can be downloaded from this link http://thesis.modusoperandi.com.au/dana.zip.

4.1 Accuracy of the Proposed Disambiguation Technique

It is labour-intensive and costly to obtain the ground truth of a dataset of such magnitude. For this experiment, we adopt the approach of finding similarities between the publications through the available abstracts. Along the line of Boyack et al. [1], we employ a combination of TF-IDF and PMRA (PubMed Related

Articles) to compute the similarity between PubMed abstracts of authors with the same names. All the stop words were firstly removed. If the similarity approaches 1, then these articles were considered to be from the same author.

The top author *"Kim Overvad"*'s publications can be taken as an example to illustrate how we evaluate the accuracy of the disambiguation algorithm. He has published 12 papers in the dataset we used. All the articles had a different affiliation but according to our algorithm the author is the same. To verify its correctness, the abstracts of all 12 articles demonstrating high level of similarity using TF-IDF and PMRA, confirming that the articles are from the same author.

Table 1(b) shows that the author disambiguation technique proposed in this paper demonstrates higher accuracy rate across the two similarity measures as ground truth, with the highest being 97.89% (929 out of 949)and the lowest being 94.9% (909 out of 949). Since PMRA has a higher coherence value [1], the string similarity measure of PMRA is considered as the accuracy of the system.

Table 1. Results and performance of our disambiguation system

Name	No.	Name	No.	Name	No.	Name	No.
Wei Wang	19	Ying Zhang	14	Li Wang	14	Wei Zhang	14
Ying Wang	13	Sang Lee	13	Ying Liu	13	Wei Li	13
Yan Li	13	Yan Wang	13	Wei Chen	12	Xiao Li	12

(a) Top 12 Ambiguious Names

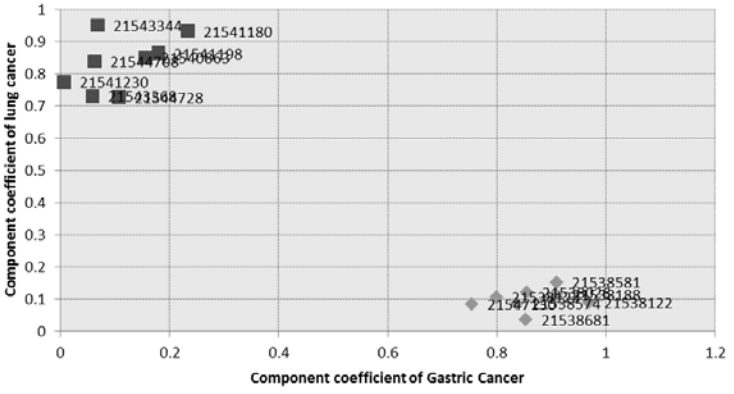
Method	No.	Accuracy
PMRA	929	97.89%
TF-IDF	909	94.94%

(b) Accuracy

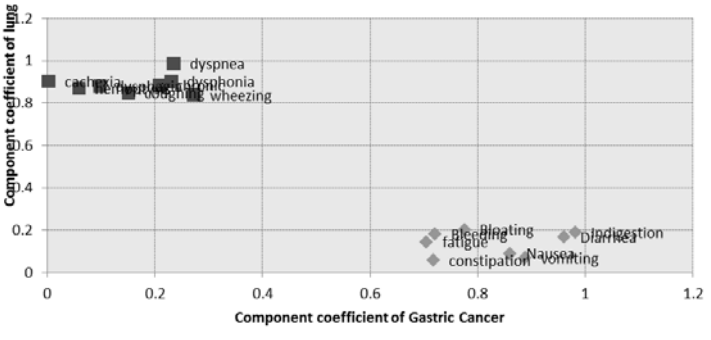
4.2 Evaluation of NetClus Results

In this evaluation, we look at the performance of NetClus on the dataset disambiguated using our technique. In Figure 2, the diamond shaped data points represent the objects in the *"Gastric Cancer"* cluster and the square shaped data points represent the objects in the *"Lung Cancer"* cluster. If the clusters are well separated, then the clustering results are considered as accurate. Partial dataset has been represented here in the scatter plots for improved readability.

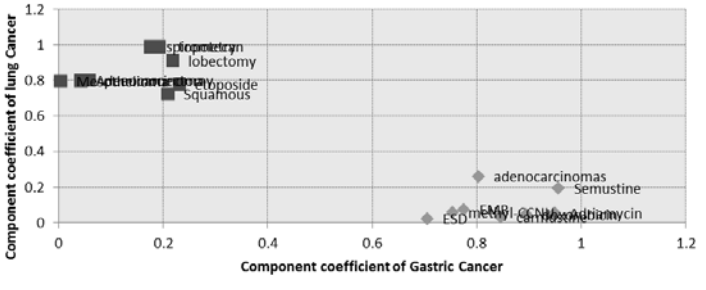
In Figure 2(a), sixteen publications were represented in this scatter plot. The PMIDs related to lung cancer are placed on the extreme left of the graph and the PMIDs related to the gastric cancer are placed on the extreme right of the graph. Both the clusters are well separated indicating the clear demarcation of data objects based on the type of cancer. In Figure 2(b), the terms are well separated based on the type of cancer. All these terms are symptoms of cancer. Symptoms of lung cancer are placed on the top left part of the chart and the symptoms of gastric cancer are placed on the bottom right of the chart. Based on the component coefficients, the objects have been placed on the chart. For example, the term *"wheezing"* is a common lung cancer symptom, therefore it is placed in the lung cancer cluster while *"constipation"* is placed in the gastric cancer cluster. In Figure 2(c), treatment and sub-types of cancer are clustered. Lobectomy refers to the surgical excision of a lobe. This may refer to a lobe of the lung. Therefore, its position in the lung cancer cluster is correct.



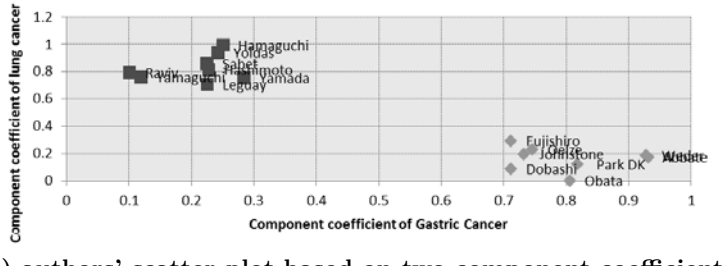
(a) PMID's scatter plot based on two component coefficients



(b) Symptoms' scatter plot based on two component coefficients



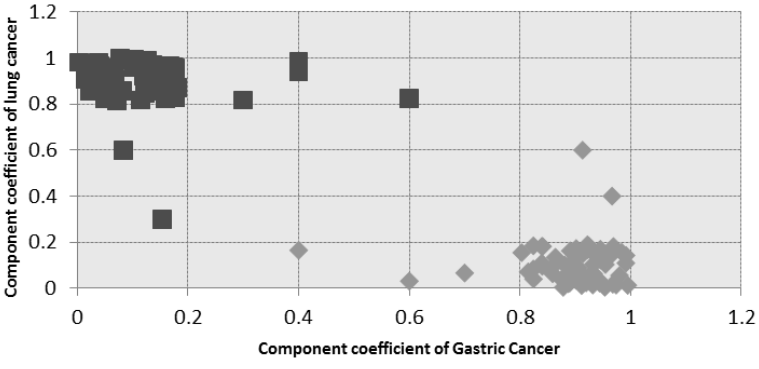
(c) Treatments' scatter plot based on two component coefficients



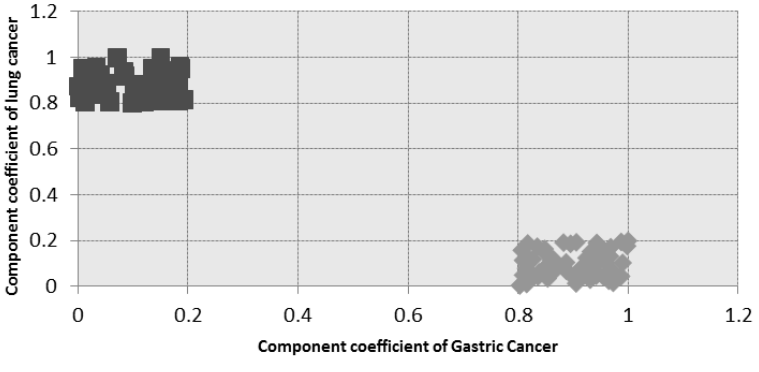
(d) authors' scatter plot based on two component coefficients

Fig. 2. Scatter Plot for articles, symptoms, treatments and authors

Similarly, since endoscopic submucosal dissection (ESD) is a mucosa resection, it is placed in gastric cancer cluster. Similarly, after author disambiguation, the authors were clustered based on the type of cancer they are researching on. The red square shaped objects represent the authors who have published more papers in lung cancer area, while the green diamond shaped objects for authors who have published more papers in gastric cancer area.



(a) Authors' scatter plot without disambiguation



(b) Authors' scatter plot with disambiguation

Fig. 3. A comparison of clustering results with/without disambiguation

Figure 3(a) and (b) compares the clustering results of authors with and without disambiguation. In Figure 3(a), few authors are scattered throughout the chart. Because the author names were ambiguous, they are not properly placed in a cluster. In Figure 3(b), authors were disambiguated using our system. Since the ambiguity was resolved, the clusters are well separated.

5 Conclusion and Future Work

To cluster datasets represented as heterogeneous information networks such as PubMed, the ambiguity amongst the objects must first be resolved. The disambiguation technique proposed in this paper was used to mitigate the ambiguity of

authors in the dataset. We have demonstrated that algorithms based on named entity recognition of affiliation string are not sufficient enough to disambiguate authors. In the proposed technique, a combination of vector space model based similarity measure, geographical information and co-author network was used to identify whether the authors with the same name are different or not.

The results of disambiguation were evaluated by comparing the PubMed abstracts of articles of disambiguated authors using text similarity algorithms such as PMRA and TF-IDF. The proposed technique showed a higher accuracy rate in both the methods. We also evaluated the NetClus algorithm with and without the disambiguation of authors. When the authors were not disambiguated, there were outliers and noise in the results. After disambiguation of authors, the clusters were clear and distinct. This disambiguation system and NetClus algorithm to identify interesting patterns that can be useful in bio-medical research. In future, we are planning to build a interface where users can use this multi-level normalisation system for author disambiguation of PubMed articles.

References

1. Boyack, K.W., Newman, D., Duhon, R.J., Klavans, R., Patek, M., Biberstine, J.R., Schijvenaars, B., Skupin, A., Ma, N., Brner, K.: Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches. PLoS ONE 6(3), e18029 (2011)
2. Golic, J.D., Mihaljevic, M.J.: A generalized correlation attack on a class of stream ciphers based on the levenshtein distance. Journal of Cryptology 3, 201–212 (1991)
3. Jonnalagadda, S., Topham, P.: Nemo: Extraction and normalization of organization names from pubmed affiliation strings. J. Biomed. Discov. Collab. 5, 50–57 (2010)
4. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Communications of ACM 18, 613–620 (1975)
5. Sun, Y., Han, J., Zhao, P., Yin, Z., Cheng, H., Wu, T.: RankClus: integrating clustering with ranking for heterogeneous information network analysis. In: EDBT 2009: Proceedings of the 12th International Conference on Extending Database Technology, pp. 565–576. ACM, New York (2009)
6. Sun, Y., Yu, Y., Han, J.: Ranking-based clustering of heterogeneous information networks with star network schema. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2009, pp. 797–806. ACM, New York (2009)
7. Torvik, V.I., Weeber, M., Swanson, D.R., Smalheiser, N.R.: A probabilistic similarity metric for medline records: A model for author name disambiguation. Journal of the American Society for Information Science and Technology 56(2), 140–158 (2005)
8. Wong, W., Liu, W., Bennamoun, M.: Integrated scoring for spelling error correction, abbreviation expansion and case restoration in dirty text. In: Proceedings of the fifth Australasian Conference on Data Mining and Analystics - AusDM 2006, vol. 61, pp. 83–89. Australian Computer Society, Inc., Darlinghurst (2006)
9. Yu, W., Yesupriya, A., Wulf, A., Qu, J., Gwinn, M., Khoury, M.: An automatic method to generate domain-specific investigator networks using pubmed abstracts. BMC Medical Informatics and Decision Making 7(1), 17–26 (2007)