Knowledge Extraction and Mining in Biomedical Research Using Rule Network Model

S.W. Chan[1], C.H.C. Leung[1], and A. Milani[2]

[1] Department of Computer Science, Hong Kong Baptist University, Hong Kong
{swchan,clement}@comp.hkbu.edu.hk
[2] Department of Mathematics & Computer Science, University of Perugia, Italy
milani@unipg.it

Abstract. Recent findings show that the quantity of published biomedical literature is increasing at a dramatic rate. Carrying out knowledge extraction from large amounts of research literature becomes a significant challenge. Here we introduce an automatic mechanism for processing such information and extracting meaningful medical knowledge from biomedical literature. Data mining and natural language processing (NLP) are applied in a novel model, called biomedical rule network model. Using this model, information and relationships among herbal materials and diseases, as well as the chemical constituents of herbs can be extracted automatically. Moreover, with the overlapping chemical constituents of herbs, alternative herbal materials can be discovered, and suggestions can be made to replace expensive treatment options with lower cost ones.

Keywords: biomedical literature, natural language processing, herb, chemical constituent, hypothesis.

1 Introduction

With rapid developments in the medical research, the quantity of published biomedical literature has been increasing dramatically in the past few decades. As shown in Fig. 1, from 1950 to 2010, the speed of publication has greatly accelerated [1], with over 2,000,000 papers published in Medline as of 2010.

Even though this may signify significant research achievements, the vast quantity of literature causes difficulty in the manual extraction of meaningful knowledge. A study [2] shows that database curators will search biomedical literature for the facts of interest, and transfer knowledge from the published papers to the database manually. It is natural that the clinicians, researchers and database curators would like to have an automatic approach to deal with the large scale data problem and discover hidden knowledge from the biomedical literature. Through the technique of Natural Language Processing (NLP), the vocabularies of biomedical literature can be extracted and classified into different classes. Recently, the focus of literature mining has been shifted from entity extraction by NLP to hidden knowledge discovery. This paper proposes a mechanism to discover the hidden relationships among the vocabularies in

K. Imamura et al. (Eds.): BHI 2013, LNAI 8211, pp. 506–515, 2013.

biomedical literature, as well as to improve the efficiency of literature analysis and hypotheses generation through Natural Language Processing and Data Mining.

2 Literature Review

Various methods have been proposed for knowledge extraction from biomedical literature, such as name of protein [30] or gene [31] extraction, protein-protein interaction [32], protein-gene interaction [33], subcellular location of protein, functionality of gene, protein synonyms [34]. In particular, [28] provides a novel approach which uses pattern discovery for knowledge extraction. The report [35] shows that several herbal medicines are identified by the U.S. Food and Drug Administration (FDA) for clinical trials for the U.S. and European markets. With increased acceptance of alternative therapies, the quantity of biomedical literature concerning Oriental Medicine is increasing. Biomedical literature mining research has shifted greater interest to Oriental Medicine literature. Another paper [26] has considered the interrelated roles of herbal materials in complex prescriptions, which utilizes data mining technique to form the association between the disease and herbal materials.

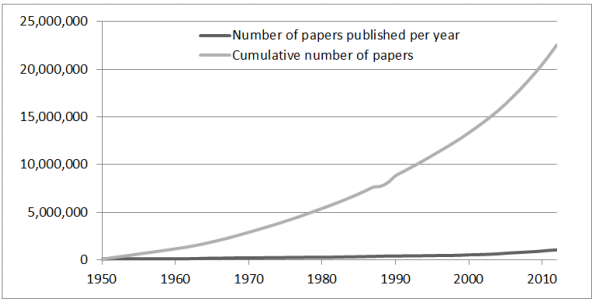

Fig. 1. The Trend of Biomedical Paper Publication in PubMed (Medline)

2.1 Natural Language Processing

NLP is able to extract information from raw texts, which can focus on sentences or vocabularies. Fig. 2 shows the text preprocess procedures before entity recognition. Raw texts are the source data which will be extracted from the database. The raw texts usually are encoded in ASCII format, which are standardized to facilitate recognition (e.g. upper case converted to lower case), while stop words, like "a, an, of, the, so, with …" are removed, and tokenization can split the text into vocabularies by space or line break or punctuation characters [4, 5]. Entities and vocabularies of interest are identified and extracted. Generally, three methodologies are implemented in entity recognition: pattern-based, dictionary-based and grammar-based. The dictionary-based methodology is the most accurate for entity extraction, but its weakness is that entities which not contained in the dictionary cannot be recognised. The pattern-based and grammar-based methodologies are relatively novel approaches which can

extract the entity without the dictionary database. However, their accuracy is not high enough because of high noise entities occurrence. In [14], a software system ABNER, has been developed for biomedical name entity recognition, with the technique of NLP, the entities are classified into five groups, protein, DNA, RNA, cell line, and cell type through different kinds of rules.

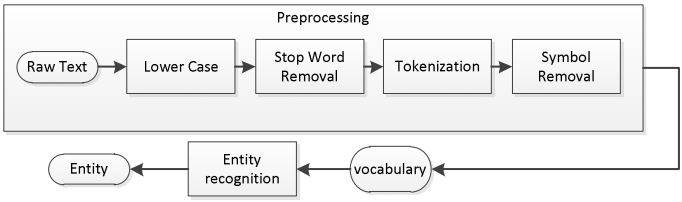

Fig. 2. The procedure of entity recognition by NLP preprocessing

2.2 Text Mining

Text mining is also referred to as text data mining. The hidden knowledge can be discovered through a large number of datasets by various data mining approaches, such as association rule, classification, clustering and so on. Another objective here is knowledge extraction, including Name Entity Recognition, Text Classification, Synonym and Abbreviation Extraction, Relationship Extraction, Integration Framework and Hypothesis Generation. The most common methods of literature mining are typically divided into several steps: text gathering, i.e. extract the raw text from the database with keyword searching; text preprocessing, i.e. convert the raw text to structured text data; data mining knowledge or module, like association rule or relationship can be formed; pruning, i.e. remove the unreasonable knowledge [3]. The accuracy and feasibility of knowledge extraction can be evaluated statistically, such as precision and recall.

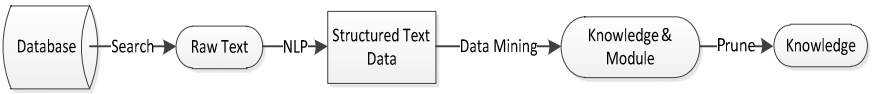

Fig. 3. The Procedures of Knowledge Extraction from Biomedical Literature

3 Biomedical Literature Entity Relation Extraction

In current biomedical literature research, information extraction is mainly focused on extracting the relationship or function of the proteins and genes. Very few research studies focus on structure analysis of Oriental Medicine, like herbal medicine of Chinese medicine. In this research, a novel model, the Biomedical Rule Network Model will be proposed. The model is able to extract the information of the relationship between the entities and generate hypotheses for future investigation. Herbal medicine is widely used in Oriental Medicine, like Traditional Chinese Medicine (TCM) and

Traditional Korean Medicine (TKM), and large amount of knowledge has been accumulated through thousands of years practice and research. However, it is not easy to understand and explain the interrelated roles of herbal material from the framework of Western Medicine, since the former has distinct concepts and unique relationships. With the structure network of chemical constituents in herbal medicine research, it can elevate the development of Oriental Medicine from their status as the collective experience of individuals into evidence-based medicine.

3.1 Data Collection and Entity Recognition

According to the latest data of Medline, there are over 22,000,000 published papers. From such a collection, we focus on the particular area of cardiovascular disease. Targeting the searching terms : "*Cardiovascular; cardiovascular disease; cardiovascular diseases; disease; drugs, chinese herbal; herb medicine, chinese traditional; pharmacognosy; Phytotherapy; plant extracts; plant preparations; plants; plants, medicinal*" 1035 results are returned. After eliminating the papers without abstract, there are 857 abstracts and these will be used as our target data. Here, the entities are the vocabularies classified into three aspects: herbal medicine, medical term and chemical constituent using the dictionary-based entity recognition technique. Various dictionaries will be referenced for entity recognition.

- Entity related to Herbal Medicine
  - Definition: The vocabulary of the herbal material name, including Latin Name, Chinese Name, Chinese Pinyin and Family Name.
  - Reference: Herbal Medicines for Human Use from European Medical Agency [16] and Medical Plant Image Database from School of Chinese Medicine in Hong Kong Baptist University [17] will be considered as reference dictionary for herb relating term extraction.
  - Examples: Abarema Clypearia (Jack) Nislsen, Cibotium Barometz, Bixa Orellena, etc.
- Entity related to Medical Term
  - Definition: The vocabulary relates to disease, diagnosis, treatment or life index
  - Reference: Unified Medical Language System (UMLS) will be the considered reference dictionary for extracting the biomedical terms. UML is a consolidated repository of medical terms and their relationships [15].
  - Examples: Hypoglycaemic Effect, Leukemia, Antioxidant Activities, Glycemic Index, Antimalarial, Vasorelaxant, Cardiovascular Diseases etc.
- Entity related to Chemical Constituents
  - Definition: The vocabulary relates to the name of chemical constituents, including trivial name and systematic name
  - Reference: Chemical Entities of Biological Interest (ChEBI) [37, 38] is a freely available dictionary accessible online
  - Examples: (Z)-3-butylidene-7-hydroxyphthalide, senkyunolide B, 3-butylphthalide, (Z)-ligustilide, etc

3.2 Association Rules

Association rules [39] have been widely used to generate relationships among entities. Here the entities of interest are herbal material, medical term and chemical constituent. The strength of an association rule is determined by the frequency of entity occurrence. An association rule can be described by an antecedent entity A, and a consequent entity B, which can be evaluated by support, confidence, and lift. For concreteness, we may take A as herbal material and B as a medical term.

$$\text{Support } (A \rightarrow B) = P(A \cap B) \tag{1}$$

Support represents the probability of herb material and medical term occurring together in dataset. The value can illustrate the popularity of the research between herbal material and material term. The confidence measures the conditional probability:

$$\text{Confidence } (A \rightarrow B) = P(B|A) \tag{2}$$

Confidence represents the credibility of the association rule between herbal material and medical term. If the value is small, this implies that among the papers that study herbal material A, there are only a few that involves medical term B. To measure the correlation between A and B, the measure lift is often used:

$$\text{Lift } (A \rightarrow B) = P(A \cap B)/[P(A) \times P(B)] \tag{3}$$

Since $P(A \cap B) = P(A) \times P(B)$ when A and B are independent, Lift $(A \rightarrow B) = 1$, if A and B are uncorrelated. Lift will be greater than one if A and B are positively correlated, and it will be less than one if A and B are negatively correlated.

3.3 Relationship Extraction

Two kinds of abstracts are contained in the dataset, validity of herbal material in particular medical usage, and the chemical constituents of herbal material. With these two types of abstracts, two types of relationships can be disclosed accordingly and integrated for hypotheses generation. In relationship extraction, it will be divided into two parts. In the first type of abstracts, association rules characterize the relationships between herbal material and medical term. An antecedent Item set A or B (herbal medicine entity) and a consequent Item C (medical term entity) (A, B $\rightarrow$ C). In the second type of abstracts, chemical constituents entity and herbal material entity will be extracted and the association rule between these entities will be formed.

In the first association network (Fig. 4), the information between herbal medicine entity and medical term entity are extracted. In some cases, more than one herb act on a particular disease or symptom. It provides the possibility that the herbs are replaceable. It is worth to know while the herbs are rare.

In the second association network (Fig. 5), the information of chemical constituents of the herb is extracted. With the combination of those two networks, the novel network of intersection of chemical constituents of herbs with particular medical usage can be formed. In the combined network (Fig. 6), herb material entities A and B are

applied for particular medical usage (medical terms), like treatment or symptom. Different herbal material entities have various chemical constituents. The hypothesis is that intersection chemical constituents of herbal material can be considered as potential effective chemical constituents for particular medical usage.

Fig. 4. Association between medical terms and the herbal material

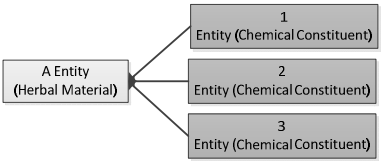

Fig. 5. Association between the herbal medicine and the chemical constituent

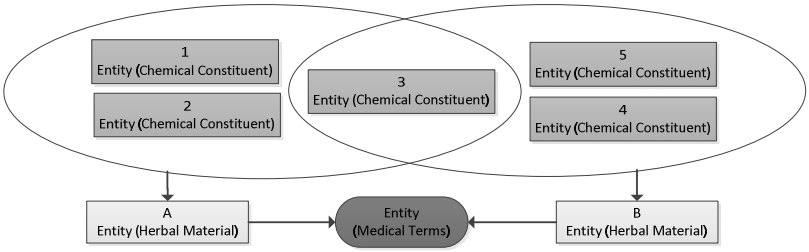

Fig. 6. Combination association network of (Herbal Material→Medical Term) and (Herbal Material→Chemical Constituent)

The interested entities are extracted from four biomedical literatures abstracts[1,2,3,4] and the hypotheses are shown in Table 1.

---

1 Kim, Eun-Young. Kim, Jung-Hyun. Rhyu, Mee-Ra.: Endothelium-Independent Vasorelaxation by Ligusticum Wallichii in Isolated Rat Aorta: Comparison of a Butanolic Fraction and Tetramethylpyrazine, the Main Active Component of Ligusticum Wallichii, 33(8), pp. 1360-1363, Biological & Pharmaceutical Bulletin (2010)

2 Wang, Jia. Yang, Jian-Bo. Wang, Ai-Guo. Ji, Teng-Fei. Su, Ya-Lun: Studies on the Chemical Constituents of Ligusticum Sinense, 34(3), pp. 378-80, English Abstract. Journal Article. Research Support, Non-U.S. Gov't (2011)

3 Matsuda, H. Murakami, T. Nishida, N. Kageura, T. Yoshikawa, M.: Medicinal Foodstuffs. XX. Vasorelaxant Active Constituents from the Roots of Angelica Furcijuga Kitagawa: Structures of Hyuganins A, B, C, and D, 48(10), pp.1429-1435, Chemical & Pharmaceutical Bulletin (2000)

4 Huang, W. H., C. Q. Song: Studies on the Chemical Constituents of Angelica Sinensis, 38(9), Yao Xue Xue Bao=Acta Pharmaceutica Sinica (2003)

Table 1. Example of information extraction

Information Extraction
1. <Herb>Ligusticum</Herb> <Medical Term>vasorelaxant</Medical Term>
2. <Herb>Ligusticum</Herb> <Chemical Constituent>levistolide A (1), (Z)-3-butylidene-7-hydroxyphthalide (2), senkyunolide B (3), 3-butylphthalide (4), (Z)-ligustilide (5), riligustilide (6), neocnidilide (7), senkyunolide A (8), beta-sitostesol (9)</Chemical Constituent>
3. <Herb>Angelica</Herb> <Medical Term>vasorelaxant</Medical Term>
4. <Herb>Angelica</Herb> <Chemical Constituent>Homosenkyunolide H (1), Homosenkyunolide I(2), Neoligustilide (3), 6-methoxycoumarin (4), Hypoxanthine-9-beta-D-ribofuranoside (5)</Chemical Constituent>

In the part of information extraction, three types of entities are extracted and they are marked respectively as <Herb>, <Medical Term> and <Chemical Constituent>. From abstract 1, the herb, ligusticum, correlates with "vasorelaxant". From abstract 2, we know that the chemical constituents listed are contained in the herb, ligusticum. The same situation can be applied in abstracts 3 and 4. According to their chemical constituents, the intersection, senkyunolide, can be found. Even though the chemical constituents are isomers (senkyunolide B, senkyunolide A, Homosenkyunolide I, Homosenkyunolide H), they may have similar effects on the human body. With above information, two hypotheses can be generated and these are shown in Table 2.

Table 2. Example of hypotheses generation

Hypotheses
– Senkyunolide might be effective chemical constituents of "vasorelaxant" that it can be found in Ligusticum and Angelica.
– The herb with the chemical constituents, senkyunolide, might be replaceable in the usage of "vasorelaxant"

3.4 Hypotheses Pruning and Evaluation

With the biomedical rule network, a number of hypotheses can be generated. However, some of the hypotheses are not worthy of investigation. For example, if the confidence of the intersection chemical constituents of hypotheses in the dataset is high, but the lift is low, it indicates that such chemical constituents are commonly contained in many herbs, but not for particular medical usage. In this case, this kind of hypotheses should not be considered further.

In the association network performance evaluation, precision and recall are able to evaluate the performance of entity recognition. Precision (4) is the fraction of retrieved entities which are relevant to the entity recognition. Recall (5) in information retrieval is the fraction of the entities that are relevant to the entity recognition.

$$precision = \frac{\{relevant\ entity\} \cap \{retrieved\ entity\}}{\{retrieved\ documents\}} \tag{4}$$

$$recall = \frac{\{relevant\ enetity\} \cap \{retrieved\ entity\}}{\{revlevant\ entity\}} \quad (5)$$

In hypotheses generation, there may already be existing research about the effective chemical constituents for particular medical usages. Those research papers can be used to validate the generated hypotheses.

4 Conclusion

The rapid growth of information in biomedical literature causes difficulties in literature review and manual analysis. In this paper, a novel analysis model of biomedical rule network is proposed, in which the techniques of natural language processing and data mining are used for hypotheses generation. Biomedical rule network of biomedical literature provides the possibility of hidden knowledge discovery between the entities of herbal material, medical term and chemical constituents. From the biomedical rule network, hypotheses that are worthy of further investigation can be generated. Promising effective chemical constituents can be mined from the intersection of herbs which has particular medical usage, such as glycemic index adjustment or vasorelaxant effect. Also, it is also possible to discover that herbs containing effective chemical constituents might be substituted by other common inexpensive herbs rather than costly rare herbs. Biomedical rule network can be applied not only to particular topic and herbal medicine, entity recognition, relationship extractions and hypotheses generation can also be applied to the other medical domain, such as AIDS treatment or drug-drug interaction. In the future, the underlying datasets can be extended to other databases, and more hypotheses can be formed with other domain datasets. With larger datasets, the performance of biomedical rule network can be improved, and the application of biomedical rule network can be usefully extended to other applicable domains.

References

1. Dan Corlan, A.: Medline Trend: Automated Yearly Statistics of PubMed Results for Any Query, `http://dan.corlan.net/medline-trend.html`
2. Berardi, M., Malerba, D., Piredda, R., Attimonelli, M., Scioscia, G., Leo, P.: Biomedical Literature Mining for Biological Databases Annotation. In: Data Mining in Medical and Biological Research, vol. 83, pp. 267–290. InTech-Open Access Publisher, University Campus STeP Ri, Slavka Krautzeka (2008)
3. Mathiak, B., Eckstein, S.: Five Steps to Text Mining in Biomedical Literature. In: Proceedings of the Second European Workshop on Data Mining and Text Mining in Bioinformatics, pp. 43–46 (2004)
4. Manning, C., Schutze, H.: Foundations of Statistical Natural Language Processing, vol. 999. MIT Press, Cambridge (1999)
5. Shatkay, H., Feldman, R.: Mining the Biomedical Literature in the Genomic Era: An Overview. Journal of Computational Biology 10(6), 821–855 (2003)
6. Cohen, A.M., Hersh, W.R.: A Survey of Current Work in Biomedical Text Mining. Briefings in Bioinformatics 6(1), 57–71 (2005)

7. Ying, L., Navathe, S.B., Civera, J., Dasigi, V., Ram, A., Ciliax, B.J., Dingledine, R.: Text Mining Biomedical Literature for Discovering Gene-to-Gene Relationships: A Comparative Study of Algorithms. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2(1), 62–76 (2005)
8. Arabie, P., Hubert, L.J.: The Bond Energy Algorithm Revisited. IEEE Transactions on Systems, Man and Cybernetics 20, 268–274 (1990)
9. Krallinger, M., Erhardt, R.A.-A., Valencia, A.: Text-Mining Approaches in Molecular Biology and Biomedicine. Drug Discover Today 10(6), 439–445 (2005)
10. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Michael Cherry, J., Davis, A.P., et al.: Gene Ontology: Tool for the Unification of Biology. Nature Genetics 25(1), 25–29 (2000)
11. Ng, S.-K., Wong, M.: Toward Routine Automatic Pathway Discovery from On-line Scientific Text Abstracts. Genome Informatics Series 10, 104–112 (1999)
12. Ono, T., Hishigaki, H., Tanigami, A., Takagi, T.: Automated Extraction of Information on Protein-Protein Interactions from the Biological Literature. Bioinformatics 17(2), 155–161 (2001)
13. Hatzivassiloglou, V., Weng, W.: Learning Anchor Verbs for Biological Interaction Patterns from Published Text Articles. International Journal of Medical Informatics 67(1), 19–32 (2002)
14. Settles, B.: An Open Source Tool for Automatically Tagging Genes, Proteins and Other Entity Names in Text. Bioinformatics 21(14), 3191–3192 (2005)
15. Unified Medical Language System - UMLS, `http://umlsks.nlm.nih.gov/`
16. Herbal Medicines for Human Use from European Medical Agency, `http://www.ema.europa.eu/ema/`
17. Medical Plant Image Database from School of Chinese Medicine in Hong Kong Baptist University, `http://library.hkbu.edu.hk/electronic/libdbs/mpd/index.html`
18. Hersh, W.: Evaluation of Biomedical Text-Mining System: Lessons Learned from Information Retrieval. Briefing in Bioinformatics 6(4), 224–256 (2005)
19. Malheiros, V., Hohn, E., Pinho, R., Mendonca, M.: A Visual Text Mining Approach for Systematic Reviews. Empirical Software Engineering and Measurement, 145–254 (2007)
20. Chapman, W.W.: Current Issues in Biomedical Text Mining and Natural Language Processing. Journal of Biomedical Informatics 42, 757–759 (2009)
21. Zuhl, M.: Automated Keyword Extraction from Bio-medical Literature with Concentration on Antibiotic Resistance. Thesis submitted to the Faculty of the Graduate School of the University of Maryland, College Park (2009)
22. Summerscales, R.L., Argamon, S., Bai, S., Huperff, J., Schwartzff, A.: Automatic Summarization of Results from Clinical Trials. Bioinformatics and Biomedicine, 372–377 (2011)
23. Berardi, M., Lapi, M., Leo, P., Loglisci, C.: Mining generalized association rules on biomedical literature. In: Ali, M., Esposito, F. (eds.) IEA/AIE 2005. LNCS (LNAI), vol. 3533, pp. 500–509. Springer, Heidelberg (2005)
24. Petrič, I., Urbančič, T., Cestnik, B.: Discovering Hidden Knowledge from Biomedical Literature. Informatica 31(1), 15–20 (2007)
25. Krallinger, M., Morgan, A., Smith, L., Leitner, F., Tanabe, L., Wilbur, J., Hirschman, L., Valencia, A.: Evaluation of Text-mining Systems for Biology: Overview of the Second BioCreative Community Challenge. Genome Biol. 9(suppl. 2) (2008)
26. Dai, H.-J., Chang, Y.-C., Tsai, R.T.-H., Hsu, W.-L.: New Challenges for Biological Text-Mining in the Next Decade. Journal of Computer Science and Technology 25(1), 169–179 (2010)

27. Chan, S.S.-K., Cheng, T.-Y., Lin, G.: Relaxation Effects of Ligustilide and Senkyunolide A, Two Main Constituents of Ligusticum Chuanxiong, in Rat Isolated Aorta. Journal of Ethnopharmacology 111(3), 677–680 (2007)
28. Bill, R.: Chinese Herbal Medicine Passes FDA Phase II Clinical Trials. HerbalEGram 7(10) (2010)
29. Hu, X., Wu, D.D.: Data Mining and Predictive Modeling of Biomolecular Network from Biomedical Literature Databases. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 251–263 (2007)
30. Fukuda, K.-I., Tsunoda, T., Tamura, A., Takagi, T.: Toward Information Extraction: Identifying Protein Names from Biological Papers. In: Proc. Pacific Symp. Biocomputing, pp. 707–718 (1998)
31. Stapley, B.J., Benoit, G.: Biobibliometrics: Information Retrieval and Visualization from Co-Occurrences of Gene Names in Medline Abstracts. In: Proc. Pacific Symp. Biocomputing, pp. 529–540 (2000)
32. Ding, J., Berleant, D., Nettleton, D., Wurtele, E.: Mining Medline: Abstracts, Sentences, or Phrases. In: Proc. Pacific Symp. Biocomputing, vol. 7, pp. 326–337 (2002)
33. Chiang, J.H., Yu, H.H.: MeKE: Discovering the Functions of Gene Products from Biomedical Literature via Sentence Alignment. Bioinformatics 19(11), 1417–1422 (2003)
34. Marcott, E.M., Xenarios, I., Eisenberg, D.: Mining Literature for Protein-Protein Interactions. Bioinformatics 17(4), 359–363 (2001)
35. Zhou, X., Peng, Y., Liu, B.: Text Mining for Traditional Chinese Medical Knowledge Discovery: A survey. Journal of Biomedical Informatics 43(4), 650–660 (2010)
36. Kang, J.H., Yang, D.H., Park, Y.B., Kimp, S.B.: A Text Mining Approach to Find Patterns Associated with Diseases and Herbal Materials in Oriental Medicine. International Journal of Information and Education Technology 2(3), 224–226 (2012)
37. Degtyarenko, K., et al.: ChEBI: A Database and Ontology for Chemical Entities of Biological Interest. Nucleic Acids Research 36(suppl. 1), D344–D350 (2008)
38. Tiago, G., Catia, P., Hugo Bastos, P.: Chemical Entity Recognition and Resolution to ChEBI. ISRN Bioinformatics (2012)
39. Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules. In: Proc. 20th Int. Conf. Very Large Data Bases, vol. 1215 (1994)