Kim et al. BMC Complementary and Alternative Medicine (2017) 17:77
DOI 10.1186/s12906-017-1596-4

BMC Complementary and Alternative Medicine

SOFTWARE Open Access

DisArticle: a web server for SVM-based discrimination of articles on traditional medicine

Sang-Kyun Kim[1*], SeJin Nam[2] and SangHyun Kim[1]

Abstract

Background: Much research has been done in Northeast Asia to show the efficacy of traditional medicine. While MEDLINE contains many biomedical articles including those on traditional medicine, it does not categorize those articles by specific research area. The aim of this study was to provide a method that searches for articles only on traditional medicine in Northeast Asia, including traditional Chinese medicine, from among the articles in MEDLINE.

Results: This research established an SVM-based classifier model to identify articles on traditional medicine. The TAK + HM classifier, trained with the features of title, abstract, keywords, herbal data, and MeSH, has a precision of 0.954 and a recall of 0.902. In particular, the feature of herbal data significantly increased the performance of the classifier. By using the TAK + HM classifier, a total of about 108,000 articles were discriminated as articles on traditional medicine from among all articles in MEDLINE. We also built a web server called DisArticle (http://informatics.kiom.re.kr/disarticle), in which users can search for the articles and obtain statistical data.

Conclusions: Because much evidence-based research on traditional medicine has been published in recent years, it has become necessary to search for articles on traditional medicine exclusively in literature databases. DisArticle can help users to search for and analyze the research trends in traditional medicine.

Keywords: Traditional medicine, Northeast Asia, MEDLINE, Support vector machine, Trend analysis

Background

MEDLINE is a bibliographic database that includes metadata and citations of biomedical literature. Although it covers articles in varying areas, including medicine, pharmacy, and biology from around the world, it does not categorize those articles by specific research area. The bibliographic content on MEDLINE is manually indexed through MeSH (Medical Subject Headings) [1], and the content can be searched for a specific topic using MeSH in Pubmed. However, because MeSH was originally designed to index, catalogue, and search articles with a controlled vocabulary thesaurus, it is difficult to apply it to the classification of academic disciplines.

Traditional medicine, particularly in Northeast Asia including traditional Chinese medicine (TCM), has developed from ancient times. A number of evidence-based articles have been published in this area in recent years. While MEDLINE also contains articles on traditional medicine, it does not offer a way to search for traditional medicine articles exclusively, making it difficult for researchers to analyze research trends in traditional medicine. Traditional medicine articles are often classified by MeSH headings such as “Medicine, Chinese Traditional”, but many articles remain without such a classification. Particularly in traditional medicine, a number of studies are being conducted in relation to herbal drugs, and these studies are generally classified by MeSH headings such as “Drugs, Chinese Herbal”. However, because studies on the effects of extracts or genomes of herbs are often identified by MeSH headings of “Plant Extracts” and “Genes, Plant”, respectively, it is not sufficient to use MeSH to determine whether the article is about traditional medicine. Therefore, in order to search for articles on traditional medicine, it is necessary to search for articles not only using MeSH, but also additional keywords. However, because different keywords

* Correspondence: skkim@kiom.re.kr
[1] Mibyeong Research Center, Korea Institute of Oriental Medicine, 1672 Yuseong-daero, Yuseong-gu, Daejeon 34054, Republic of Korea
Full list of author information is available at the end of the article

will bring different search results, it is difficult to search exclusively for traditional medicine articles.

In academic disciplines, there generally exist journals that mainly publish articles for a specific discipline. However, all of the articles in the discipline are not always published in the given journal and literature databases such as MEDLINE include many journals covering various areas. Therefore, it is difficult to discriminate articles on traditional medicine from those of other disciplines by using only the journal information.

In order to overcome these difficulties, this research devised a classifier to identify articles on Northeast Asian traditional medicine by using the Support Vector Machine (SVM), which is widely used in text mining. We also constructed a web server called DisArticle, in which only articles on traditional medicine can be searched for from among all articles in MEDLINE. The major goal of DisArticle was to reduce the workload of researchers by reducing the number of articles they search and identify. This can help them to easily analyze research trends in traditional medicine.

Much research on machine learning techniques has been done, such as classification based on the MEDLINE database. The research on classification mainly has been done to discover new knowledge such as protein-protein interactions [2] or gene disease associations [3]. This research has been also used to extract gene terms [4] or chemical names [5] within the content of articles. Recently, an SVM-based classifier was constructed to determine whether a certain article describes a randomized clinical trial (RCT) [6]. MEDLINE not only includes the article publication type of the RCT, but also defines what work is about the RCT (http://www.ncbi.nlm.nih.gov/mesh/68016449). However, because the identification of RCTs is conducted in a simple way, this study proposes a classifier model to identify RCT articles using only the metadata and MeSH terms of each article.

Implementation

Data preparation

The corpus used to establish a classifier to identify traditional medicine articles was prepared directly from MEDLINE, with PubMed identifiers (PMID) ranging from 22,000,000 to 22,240,000. To increase the recall of the classifier, we only selected articles with both a title and an abstract, leading to a total of 189,674 articles. After the texts were extracted from the titles, abstracts, keywords, affiliation, journal name, and MeSH, they were indexed using Apache Lucene [7] to search for the features of the classifier. Two researchers in traditional medicine then manually reviewed the articles to determine whether they were about traditional medicine. The classification criteria were the biomedical research in relation to the terms within the WHO International Standard Terminologies (IST) [8], or medical herbs listed in the pharmacopoeia of China, Korea, and Japan. The WHO IST was constructed to provide a standardized nomenclature of traditional medicine in the Western Pacific region including China, Korean, and Japan. The nomenclature includes technical terms such as theories, diagnostics, therapeutics, acupuncture, and moxibustion. Because WHO IST excludes herbal drugs, we extracted the names of herbs from the Chinese, Korean, and Japanese government-published pharmacopeias. Cohen's kappa for inter-rater agreement was 0.96, which is considered as substantial agreement. The two reviewers finally selected 1,537 traditional medicine articles after reaching a consensus. This paper used those articles as a gold standard for our classifier.

Training and testing methods

This research used WEKA [9] to identify articles on traditional medicine in the MEDLINE database. A Java language-based software developed by the University of Waikato, WEKA provides not only a range of machine learning algorithms for data mining, but also a number of tools for data pre-processing, classification, and visualization. In particular, WEKA implements the Sequential Minimal Optimization (SMO) algorithm [10] to train the SVM classifier.

The number of articles on traditional medicine is very small compared to the total number of articles in MEDLINE. Thus, for the training set used for the classifier model, gold standard articles and additional articles in quantity nine times the number of gold standard articles were randomly selected from the PMID range mentioned above, producing a total of 15,370 articles. The features for the training of the classifier included the title, abstract, keywords, affiliation, journal title, MeSH, and herbs. The herbal data consisted of the Latin, common, and scientific names of herbs in the Korean, Chinese, and Japanese pharmacopoeia as well as the medicinal parts of these herbs. The herbal feature is a binary value to determine whether such data exist in the title and abstract of each article. For the features other than herbs, the StringToWordVector filter provided by WEKA was used to create a vector for the words frequently found in each feature. The SMO algorithm used attributes, consisting of both the binary value of the herbs and the word vector as the kernel input, and finally classified all article instances as 1 or 0. The SVM classifier was trained with a polykernel with the default parameters provided in Weka and was tested with 10-fold cross validation. Figure 1 shows the overall process for the implementation of our web server, in which the stages of the solid line boxes are described in the results section.

Performance measure

The objective of our classifier is to discriminate articles on traditional medicine from those of other disciplines.

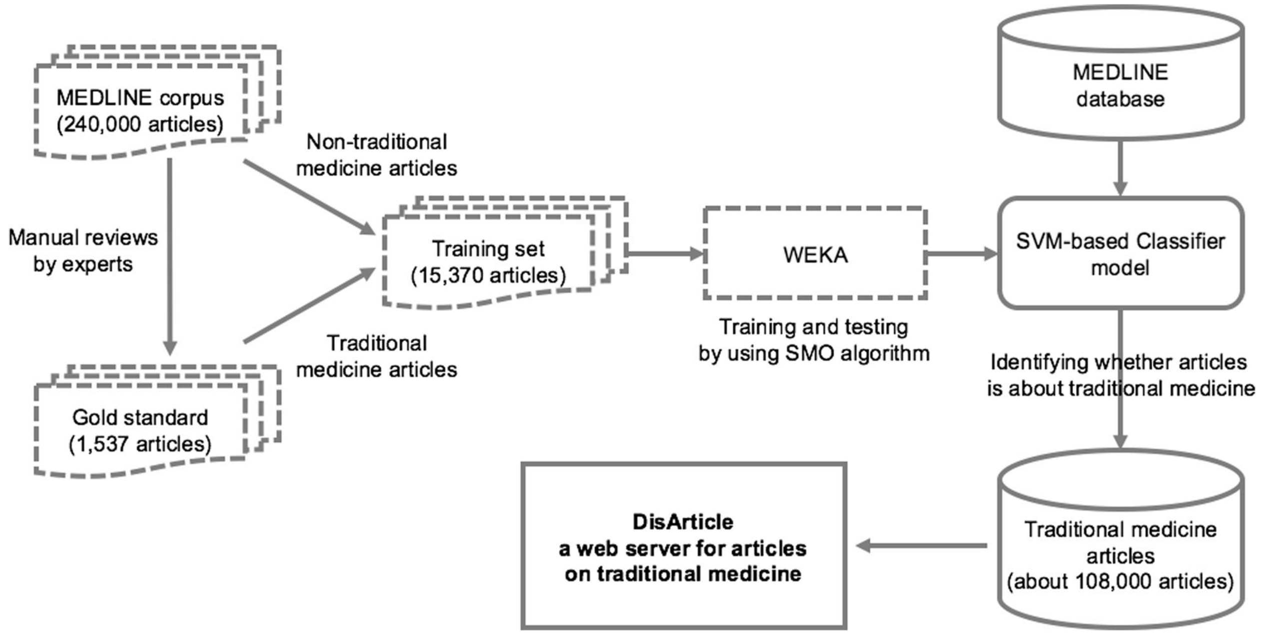


Fig. 1 Process of implementation of our web server

Using the articles on traditional medicine as the gold standard, the terminology used to evaluate of our classifier are follows:

- True positive (TP) – an article is identified as a traditional medicine article by both the classifier and the gold standard.
- True negative (TN) – an article is not identified as a traditional medicine article by either the classifier or the gold standard.
- False positive (FP) – an article is identified as a traditional medicine article by the classifier but not by the gold standard.
- False negative (FN) – an article is identified as a traditional medicine article by the gold standard but not by the classifier.

Precision, recall, and F-measure are calculated in the usual way:

- Precision = TP/(TP + FP)
- Recall = TP/(TP + FN)
- F-measure = (2 * Precision * Recall)/(Precision + Recall)
- Accuracy = (TP + TN)/(TP + TN + FP + FN)

Results

Classifier model

Because our classifier uses many attributes to identify traditional medicine articles, the number of attributes affects the classifier performance. Table 1 shows the performance of identifying traditional medicine articles with varying numbers of attributes created for all features. In the results, the ALL-140 classifier model with 140 attributes was observed to have the highest F-measure, and the level of performance dropped when the number of attributes was too big or small. The list of attributes and the detailed accuracy of the ALL-140 classifier model are given in the Additional file 1.

Table 2 shows the performance of the classifier models with different combinations of features. The number of attributes was maintained at 140, which value showed the highest performance, as can be seen in Table 1. The ALL-140 classifier had the best performance with an F-measure of 0.93. While the best precision obtained was 0.957 when using the TAK + H classifier excluding affiliation, journal title, and MeSH, the recall of the classifier was relatively small compared to ones that used herbal data (H). In general, a classifier with more

Table 1 Performance of classifiers by number of attributes (The value of the number in the name of the classifiers indicates the number of attributes used in that classifier)

	Precision	Recall	F-Measure	Accuracy
ALL-40	0.927	0.886	0.906	0.982
ALL-90	0.955	0.88	0.916	0.984
ALL-140	0.952	0.91	0.93	0.986
ALL-190	0.951	0.903	0.926	0.986
ALL-240	0.949	0.904	0.926	0.985
ALL-290	0.948	0.899	0.923	0.985
ALL-340	0.945	0.902	0.923	0.985

Table 2 Performance of the classifiers by feature (ALL: all attributes, TAK: title, abstract, keyword, H: herbs, A: affiliation, J: journal, M: MeSH)

	Precision	Recall	F-Measure	Accuracy
TAK	0.858	0.593	0.702	0.950
TAK + M	0.887	0.682	0.771	0.959
TAK + AJM	0.879	0.71	0.785	0.961
TAK + H	0.957	0.862	0.907	0.982
TAK + HA	0.953	0.889	0.92	0.985
TAK + HJ	0.954	0.893	0.923	0.985
TAK + HM	0.954	0.902	0.927	0.986
TAK + HAJ	0.95	0.897	0.923	0.985
TAK + HAM	0.95	0.908	0.928	0.986
TAK + HJM	0.952	0.903	0.927	0.986
ALL-140	0.952	0.91	0.93	0.986

features tended to show better performance. In particular, when the feature of herbal data was added, the performance increased significantly, which indicates that herbal data is an important feature in identifying articles on traditional medicine. While the feature of MeSH also increases the performance, herbal data is a more important feature than MeSH in the performance. It was shown that the performance of the TAK-AJM classifier excluding only herbal data is lower than that of TAK-H using only herbal data. For this reason, we do not include the performance of all combinations of features without herbal data in Table 2.

Web server

Articles on traditional medicine were discriminated from among those registered on MEDLINE with titles and abstracts up to January 8, 2016. Using the TAK + HM classifier model, a total of about 108,000 articles were identified as being about traditional medicine. Although the model test showed the best F-measure for the ALL-140, this result was not significantly better than that of the TAK + HM. Moreover, because precision is generally more important than recall in a web server used for searching [11], we chose the TAK + HM classifier model, which has higher precision, for our web server. The highest precision was found with the TAK + H; however, we did not choose this classifier because its recall is relatively small. The list of attributes and the detailed accuracy of the TAK + HM classifier model are provided in the Additional file 1.

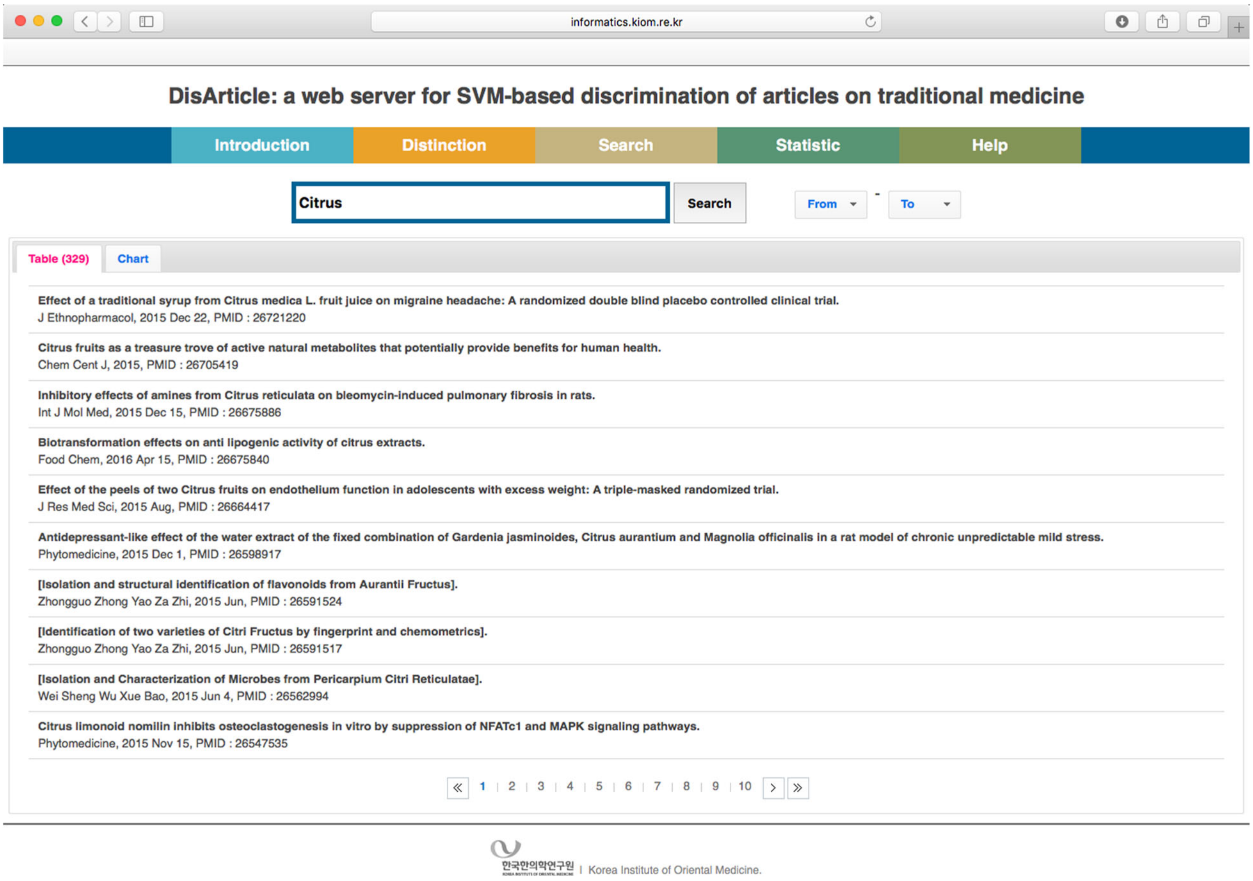
With the article database established in this manner, this study built a web server called DisArticle in which users can search for articles on traditional medicine. DisArticle also enables users to identify other articles besides those in MEDLINE and to obtain statistical data such as the top ten researched herbs and the top ten journals publishing the most articles on traditional medicine. All these functions are provided at <http://informatics.kiom.re.kr/disarticle>. The main menu of this web server consists of Distinction, Search, and Statistics.

The Distinction menu enables users to identify whether an article is on traditional medicine. When a user submits the PMID of a certain article or its metadata, the distinction result is shown. Because distinction is also possible with the metadata of an article, any article that has not been identified by our web server can be identified in this menu. The web server also provides the identification API, so that a number of articles can be identified at the same time. In the Search menu, users can search for articles already identified as traditional medicine articles in our web server. When the keyword is inputted into the search field and the "Search" button is clicked, a list of articles matching the keyword is shown. In order to provide the annual distribution, the search results are also given as a bar chart, in which the x axis is the year and y axis the number of articles. The Statistics menu currently provides two sorts of statistical data about articles on traditional medicine. One is the annual distribution of articles on the top ten researched herbs among all traditional medicine articles, and the other is the annual distribution of the top ten journals that publish the most articles on traditional medicine. A detailed explanation of the use of our web server is given in the Help menu, and a full PMID list of the gold standard used in our classifier can be downloaded from the Introduction menu of our web server.

Figure 2 shows an example of a search for articles with respect to *Citrus* species in our web server. *Citrus unshiu* Markovich and *C. reticulata* Blanco are *Citrus* species known as "Jinpi" (Korean) or "Chenpi" (Chinese). In Northeast Asia including China, Korea, and Japan, the dried peels of these species have been used since ancient times as traditional herbal drugs for the treatment of gastrointestinal and inflammatory diseases [12–14]. Recently, as the peeled fruit is known as a rich source of flavonoids and carotenoids, several studies have been done with respect to the edible tissue of citrus fruits [15]. Therefore, if the keyword "Citrus" is searched in Pubmed, many articles with respect to the whole fruit including the peel and the pulp of *C. unshiu* Markovich and *C. reticulata* Blanco are listed. However, our web server can show only research on citrus fruits used in traditional herbal drugs.

Discussion

The efficacy of traditional medicines from Northeast Asia, including TCM, has been proven clinically since ancient times. With the development of modern medicine, a number of published studies have shown



DisArticle: a web server for SVM-based discrimination of articles on traditional medicine

Effect of a traditional syrup from Citrus medica L. fruit juice on migraine headache: A randomized double blind placebo controlled clinical trial.
J Ethnopharmacol, 2015 Dec 22, PMID : 26721220

Citrus fruits as a treasure trove of active natural metabolites that potentially provide benefits for human health.
Chem Cent J, 2015, PMID : 26705419

Inhibitory effects of amines from Citrus reticulata on bleomycin-induced pulmonary fibrosis in rats.
Int J Mol Med, 2015 Dec 15, PMID : 26675886

Biotransformation effects on anti lipogenic activity of citrus extracts.
Food Chem, 2016 Apr 15, PMID : 26675840

Effect of the peels of two Citrus fruits on endothelium function in adolescents with excess weight: A triple-masked randomized trial.
J Res Med Sci, 2015 Aug, PMID : 26664417

Antidepressant-like effect of the water extract of the fixed combination of Gardenia jasminoides, Citrus aurantium and Magnolia officinalis in a rat model of chronic unpredictable mild stress.
Phytomedicine, 2015 Dec 1, PMID : 26598917

[Isolation and structural identification of flavonoids from Aurantii Fructus].
Zhongguo Zhong Yao Za Zhi, 2015 Jun, PMID : 26591524

[Identification of two varieties of Citri Fructus by fingerprint and chemometrics].
Zhongguo Zhong Yao Za Zhi, 2015 Jun, PMID : 26591517

[Isolation and Characterization of Microbes from Pericarpium Citri Reticulatae].
Wei Sheng Wu Xue Bao, 2015 Jun 4, PMID : 26562994

Citrus limonoid nomilin inhibits osteoclastogenesis in vitro by suppression of NFATc1 and MAPK signaling pathways.
Phytomedicine, 2015 Nov 15, PMID : 26547535

Fig. 2 Result of search for the keyword “Citrus”

scientific evidence on the efficacy of traditional medicine. What significantly distinguishes traditional medicine in Northeast Asia from other regional traditional medicines, such as traditional African medicine or Ayurveda medicine, is mainly the use of herbal drugs [16]. Certain medicinal herbs are only grown in Northeast Asia, and the same herbs may produce different medicinal effects depending on the region where they are grown. This research showed that herbal data are an important feature for identifying articles on traditional medicine. Therefore, our classifier should be trained with herbal data from other regions to provide broader coverage of other traditional medicines from around the world.

Recently, because of the increasing global interest in healthcare, a number of studies on complementary and alternative medicines (CAMs) are being done. CAMs are generally known as any medicinal practice that does not originate from scientific evidence. The CAM category includes TCM and other herbal medicines, as well as non-traditional medicines [17]. In order to build a web server to identify articles on CAMs, it is of foremost necessity to define criteria that can be used to show whether a certain article is about a CAM. This study used the WHO IST and the pharmacopoeia of Korea, China, and Japan as the criteria to identify articles on Northeast Asian traditional medicine. If the criteria for the identification of CAM articles is defined, it will be possible to establish an article identification system for CAMs similar to our web server.

As stated in the background, the work of Cohen et al. is similar to ours. We both constructed SVM-based binary classifiers for articles on MEDLINE and provided methods to determine whether an article is in a particular research area or of a particular type. Cohen et al. used only metadata and MeSH to identify articles. However, to achieve good performance, identifying articles in the traditional medicine field requires more features such as those that describe medicinal herbs.

In future work, we will do experiments with not only the SVM but also with a variety of machine learning algorithms. In addition, although the web server currently provides only basic statistical data on articles on traditional medicine, it will be updated to provide more professional trend analysis [18] or meta-analysis [19] data about traditional medicine articles.

Conclusions

DisArticle provides a method to search for articles on Northeast Asian traditional medicine. In order to construct DisArticle, we first devised an SVM-based classifier model to identify articles on traditional medicine. The classifier model was made to discriminate between articles on traditional medicine and those of other disciplines for all articles that have both titles and abstracts in MEDLINE. DisArticle makes it possible for researchers to search for articles on traditional medicine, helping them to understand annual trends in traditional medicine research.

Additional file

Additional file 1: Information on the ALL-140 and TAK-HM classifier model. (DOCX 97 kb)

Abbreviations

CAM: Complementary and alternative medicine; DisArticle: Discrimination of articles; MeSH: Medical subject headings; PMID: PubMed identifier; RCT: Randomized clinical trial; SMO: Sequential minimal optimization; SVM: support vector machine; TCM: traditional Chinese medicine; WHO IST: World Health Organization International Standard Terminologies

Funding

This research was supported by a National Research Foundation (NRF) grant funded by the Korean government, MSIP (grant number 2015R1C1A1A01052309 and 2015M3A9E3051024), and the Korea Institute of Oriental Medicine (grant number K16512).

Availability of data and materials

Project name: DisArticle, Project home page: http://informatics.kiom.re.kr/disarticle, Operating system: Platform independent, Programming language: Java, Other requirement: Java 1.6 or higher, License: GNU GPL, Any restrictions to use by non-academics: none.

Authors' contributions

SKK, SN and SK conceived the idea. SKK and SN participated in the data mining and the system implementation. SKK and SK reviewed the articles. SKK wrote the paper and all authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Author details

[1]Mibyeong Research Center, Korea Institute of Oriental Medicine, 1672 Yuseong-daero, Yuseong-gu, Daejeon 34054, Republic of Korea. [2]National Center of Excellence in Software, Chungnam National University, 99 Daehak-ro, Yuseong-gu, Daejeon 34134, Republic of Korea.

Received: 13 May 2016 Accepted: 20 January 2017
Published online: 28 January 2017

References

1. Bachrach CA, Charen T. Selection of MEDLINE contents, the development of its thesaurus, and the indexing process. Med Inform (Lond). 1978;3(3):237–54.
2. Marcotte EM, Xenarios I, Eisenberg D. Mining literature for protein-protein interactions. Bioinformatics. 2001;17(4):359–63.
3. Ananiadou S, Kell DB, Tsujii J. Text mining and its potential applications in systems biology. Trends Biotechnol. 2006;24(12):571–9.
4. Yu H, Agichtein E. Extracting synonymous gene and protein terms from biological literature. Bioinformatics. 2003;19 Suppl 1:i340–9.
5. Zhang JD, Geer LY, Bolton EE, Bryant SH. Automated annotation of chemical names in the literature with tunable accuracy. J Cheminform. 2011;3(1):52.
6. Cohen AM, Smalheiser NR, McDonagh MS, Yu C, Adams CE, Davis JM, Yu PS. Automated confidence ranked classification of randomized controlled trial articles: an aid to evidence-based medicine. J Am Med Inform Assoc. 2015;22(3):707–17.
7. McCandless M, Hatcher E, Gospodnetić O, Gospodnetić O. Lucene in action. 2nd ed. Manning: Greenwich; 2010.
8. Pacific WHOROftW. WHO international standard terminologies on traditional medicine in the Western Pacific Region. Manila: World Health Organization, Western Pacific Region; 2007.
9. Frank E, Hall M, Trigg L, Holmes G, Witten IH. Data mining in bioinformatics using Weka. Bioinformatics. 2004;20(15):2479–81.
10. Platt J. Fast Training of Support Vector Machines using Sequential Minimal Optimization. In: Advances in Kernel Methods. MIT Press; 1999. p.185-208. http://dl.acm.org/citation.cfm?id=299105.
11. Van Rijsbergen CJ. Information retrieval. 2nd ed. London; Boston: Butterworths; 1979.
12. Noh HJ, Hwang D, Lee ES, Hyun JW, Yi PH, Kim GS, Lee SE, Pang C, Park YJ, Chung KH, Kim GD, Kim KH. Anti-inflammatory activity of a new cyclic peptide, citrusin XI, isolated from the fruits of Citrus unshiu. J Ethnopharmacol. 2015;163:106–12.
13. Kim HG, Kim GS, Lee JH, Park S, Jeong WY, Kim YH, Kim JH, Kim ST, Cho YA, Lee WS, Lee SJ, Jin JS, Shin SC. Determination of the change of flavonoid components as the defence materials of Citrus unshiu Marc. fruit peel against Penicillium digitatum by liquid chromatography coupled with tandem mass spectrometry. Food Chem. 2011;128(1):49–54.
14. Lyu JH, Lee HT. Effects of dried citrus unshiu peels on gastrointestinal motility in rodents. Arch Pharm Res. 2013;36(5):641–8.
15. Nogata Y, Sakamoto K, Shiratsuchi H, Ishii T, Yano M, Ohta H. Flavonoid composition of fruit tissues of citrus species. Biosci Biotechnol Biochem. 2006;70(1):178–92.
16. Patwardhan B, Warude D, Pushpangadan P, Bhatt N. Ayurveda and traditional Chinese medicine: a comparative overview. Evid Based Complement Alternat Med. 2005;2(4):465–73.
17. Cheung F. TCM Made in China. Nature. 2011;480(7378):S82–3.
18. Deshazo JP, Lavallie DL, Wolf FM. Publication trends in the medical informatics literature: 20 years of "Medical Informatics" in MeSH. BMC Med Inform Decis Mak. 2009;9:7.
19. Haidich AB. Meta-analysis in medical research. Hippokratia. 2010;14 Suppl 1:29–37.