Association for Information Systems
AIS Electronic Library (AISeL)

PACIS 2018 Proceedings

Pacific Asia Conference on Information Systems (PACIS)

6-26-2018

An Integrated Web-based System for MEDLINE Analysis: A Case Study of Chronic Kidney Disease

Yi-Ling Lin
Department of Management Information Systems, yl_lin@nccu.edu.tw

Wei-En Huang
Department of Information Management, National Sun Yat-Sen University, weienhuang13@gmail.com

Peir-In Liang
Department of Pathology, Kaohsiung Medical University Hospital, Kaohsiung Medical University, peirinl@yahoo.com

Chun-Wei Tung
School of Pharmacy, Kaohsiung Medical University, cwtung@kmu.edu.tw

An Integrated Web-based System for MEDLINE Analysis: A Case Study of Chronic Kidney Disease

Completed Research Paper

Yi-Ling Lin
Department of Management Information Systems, National Chengchi University
Taipei, Taiwan
yl_lin@nccu.edu.tw

Wei-En Huang
Department of Information Management, National Sun Yat-sen University
Kaohsiung, Taiwan
weienhuang13@gmail.com

Peir-In Liang
Department of Pathology, Kaohsiung Medical University Hospital, Kaohsiung Medical University
Kaohsiung, Taiwan
peirinl@yahoo.com

Chun-Wei Tung
School of Pharmacy, Kaohsiung Medical University
Kaohsiung, Taiwan
cwtung@kmu.edu.tw

Abstract

In the era of big data, medical researchers attempt to utilize some analysis techniques like machine learning and text mining on their large-scale corpora to save valuable labor work and time. Consequently, many data analysis platforms are built to support medical professionals such as Pubtator, GeneWays, BioContext, etc. These platforms are helpful to medical entities recognition and relation extraction, but there is not an integrated platform to support researchers' various needs, and medical projects are isolated from each other, which is hard to be shared and reused. As a result, we present an integrated system containing 'name entity recognition', 'document categorization' and 'association extraction'. Besides, we add the concept of 'socialization' making projects reusable for further analyses. A case study of chronic kidney disease was adopted to indicate the effectiveness of the proposed system.

Keywords: Text mining, machine learning, name entity recognition, association extraction, medical analysis, sharing

Introduction

Along with Internet diffusion, the amount of the medical literature grows rapidly. It is no longer easy and fast to retrieve medical data without performing a time-consuming search task. To efficiently dig in the numerous medical documents online, many researchers attempted to adopt text mining and machine learning techniques such as document categorization, named entity recognition, natural language processing, and association recognition (Cooper and Kershenbaum 2005; Coulet et al. 2010; Papanikolaou et al. 2015; Stephens et al. 2001). However, data in medical research is complicated to retrieve for different professionals' needs by performing sophisticated search and processing to reach their searching goals. Although text mining and machine learning techniques have been wildly used in the medical domain, most of the medical professionals are not familiar with these techniques. There is

a convenient application for doing machine learning, Weka[1], but medical professionals still need to spend their time on familiarizing themselves with this system and the analysis models. It is even hard for them to conduct any text mining techniques on their own. Although general data analysts are good with those text mining techniques, as we all know that the conducing a medical related analysis requires high level of domain knowledge.

In general, there are iteratively communications between medical professionals and data analysts when medical professionals need to utilize some techniques of text mining and machine learning. Under this circumstance, some medical support systems based on text mining techniques are developed to support domain experts on medical studies. For example, Pubtator (Wei et al. 2013) has a comprehensive and efficient name entity recognition function for assisting curation on PubMed abstracts. DISEASES (Pletscher-Frankild et al. 2015) uses text mining and integrates diverse databases to assist in disease-gene associations searching. PolySearch2 (Liu et al. 2015) is developed for identifying relationships of many biomedical entities. These platforms and systems have essential functions for supporting different medical purposes. However, they are disassembled and decentralized, which makes researchers have to get different functions from different platforms. Besides, since the curated results from each different research are scattered, researchers have to spend time and labor on curating similar datasets for the same purpose frequently.

As a result, this study aims to reduce the barrier and inconvenience of large-scale medical data analysis by developing an integrated web-based medical analysis platform which incorporates state-of-the-art techniques in text mining and machine learning to analyze MEDLINE abstracts. Meanwhile, the system also aims to become a social community platform allowing people with similar interests to access and expand similar curated datasets. The proposed platform contains three major functions including "document categorization", "relation extraction", and "socialization". By practically applying the proposed system with a real chronic kidney toxicity case, the system shows its effectiveness and usefulness on facilitating professionals to categorize the publications of kidney related diseases and explore associations between the disease entities and the compound entities.

Related Work

Text mining in medical studies

Recent years, researchers have used text mining approaches to retrieve information from medical documents. Large amounts of studies are interested in mining relations and interactions between some entities such as diseases, genes, proteins, drugs, etc. Stephens et al. (2001) applied natural language processing method and co-occurrence method to discovery relations of genes from MEDLINE abstracts. Cooper and Kershenbaum (2005) utilized text analytics, statistical and graphical analysis, and a set of easily implemented rules to detect interactions between some proteins in MEDLINE abstracts. Coulet et al. (2010) adopted name entity recognition techniques to parse PGx entities such as genes, drugs, and phenotypes (e.g., VKORC1, warfarin, clotting disorder) from MEDLINE abstracts and extracts commonly occurring PGx relationships. Papanikolaou (2015) focused on text mining-based computational methodologies to explore proteins and the corresponding interactions from biological literature and databases. It shows that text mining techniques can effectively support researchers to extract information from medical documents.

Systems for supporting medical analysis

Several systems are developed for assisting medical researchers to use text mining techniques in their studies. Rzhetsky et al. (2004) developed GeneWays to analyze interactions between molecular substances and to visualize molecular networks. Gerner et al. (2012) presented an integrated text mining system, BioContext, which extracts, extends and integrates results from several text mining tools for entity recognition and event extraction. Stenetorp et al. (2012) developed a web-based annotation tool,

[1] https://www.cs.waikato.ac.nz/ml/weka/

BRAT, which implements natural language processing techniques to support manual annotation efforts and to increase annotation efficiency. Bravo et al. (2015) built another text mining system, BeFree, which identifies gene-disease, drug-disease, and drug-target associations from the manual-curated document set. Although these systems indicate that it is helpful to medical professionals in analyzing medical documents by developing a text mining platform, they all ignore some extended functions to make the medical analysis more completely. GeneWays has good performance on extracting gene relation and visualizing the result, but it focuses on gene only and cannot identify the different types of relations. BioContext is helpful to recognize gene/protein, but it does not use its good performance on NER to do further relation extraction. BRAT is useful for automatically annotating entities and dependency analysis, but it does not sort out the relationship between different entities. As a result, we dedicated to building an integrated platform to make the medical analysis more convenient and comprehensive.

Material and Methods

In this study, we developed a platform that integrates functions as following: *"name entity recognition"* adopting dictionary-based method and applying Pubtator's API to support researchers on observing the entities in documents, *"document categorization"* grouping documents into different categories and helping researchers focus on the groups of documents they are interested in, *"association extraction"* discovering relationships between different entities, and *"socialization"* helping researchers share their results to others who are interested in the same direction and attempting to refer or extend them. Currently, this study focuses on the first three text mining related parts, and the socialization is still under developing. The system diagram is shown in Figure 1.

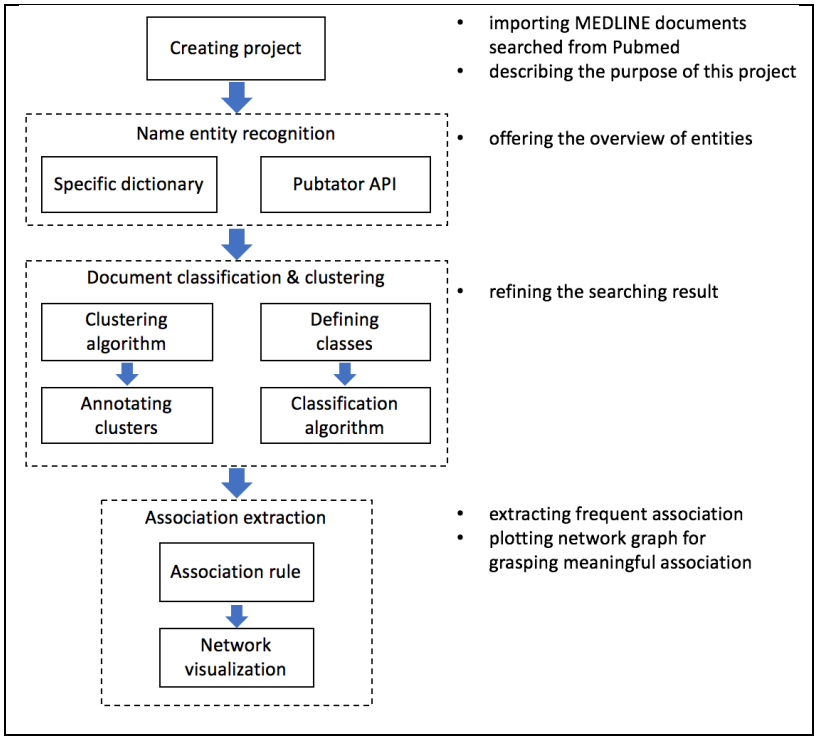

Figure 1. System diagram

Name entity recognition

The popular methods of named entity recognition contain dictionary-based methods, rule-based methods, and machine learning methods (Zhu et al. 2013). The dictionary-based method is the most direct and intuitive approach. However, this method requires complete dictionary support, which is known to be difficult to recognize the up-to-date words. Comparing to the dictionary-based method, the rule-based method is relatively easy on updating trending words, but the completeness of collected rules

is the essential factor for the value of the method. As a result, the rule-based method is known to be less flexible to fit various conditions. Recently, some researchers adopt the machine learning method in biomedical domain. He et al. (2014) combined conditional random fields (CRF) model and dictionary look-up to recognize drug names and Kazama et al. (2001) and Neelakantan and Collins (2014) applied Support Vector Machines (SVMs) to biomedical named entity recognition. Although the machine learning performs well, the technique usually needs a comprehensive training model to enhance its performance, and a comprehensive training model is usually based on a complete dictionary. Since the medical domain has many existing dictionaries and it is difficult to make all possible rules, this study adopted the dictionary-based approach as the initial test bed.

In our system, we adopted several common-used dictionaries from two trustworthy platforms, PubMed Resources[2] and the Comparative Toxicogenomics Database[3](CTD). *PubMed Resources* provides an easy FTP download mechanism and contains databases including genes, chemicals, proteins, etc. *CTD* contains entities including genes, chemicals, and diseases, and besides, it provides wide-ranging synonyms and relationships between these entities. These two sources comprise 164,639 chemical words and 621 kidney related diseases and persistently update their databases, so we can lower the probability of missing up-to-date words by including these two dictionaries into our system. In addition to these common-used dictionaries, our system allows scholars to import other vocabularies, and we also implemented Pubtator's API to support our dictionary-based method.

Document categorization

There are two major approaches to split corpora into different groups of documents which are unsupervised- and supervised-learning. Unsupervised-learning, called clustering, is an approach which can rapidly group corpora into *k* clusters and does not need to annotate labels first. The most commonly used unsupervised-learning approach is the K-means algorithm (Huang 2008), and recent studies have adopted topic modeling approach such as LDA to group corpora (Xie and Xing 2013). However, the unsupervised-learning approach needs to determine a number of clusters first, and a commonality between the documents in each cluster needs to be interpreted by professionals. In contrast to the unsupervised-learning approach, the supervised-learning approach, classification, can categorize corpora into different classes which are defined by the researchers, so researchers can know the difference between classes. However, the supervised-learning model needs a great amount of ground truth documents for its learning, and usually the ground truth documents require professionals to annotate, which is a time- and labor-consuming task. In our system, these two approaches are provided. Researchers can do the unsupervised-learning approach first to initially grasp how these documents can be grouped. If the performance of the unsupervised-learning approach is not very well and researchers cannot easily to distinguish the difference between these clusters, researchers could choose to define the specific categories and then curate some ground truth documents to train a supervised-learning model.

Preprocessing

To avoid meaningless words (i.e. stop words) like conjunction (e.g. 'or' and 'and') or pronounce such as 'it', and 'he', we applied NLTK's (Natural Language Toolkit) common stop-word list to clear them. In this way, we kept most of the useful words as features to promote the categorization performance. After removing the stop words, same words in different inflected forms were also normalized. Lemmatization function (Björkelund et al. 2010) is implemented in our system. Unlike stemming process removes the affixes without knowledge, lemmatization considers the part-of-speech. Therefore, when a word has different meanings depending on the part-of-speech, lemmatization can more effectively return a word to its original form.

[2] https://www.ncbi.nlm.nih.gov/guide/all/

[3] http://ctdbase.org/

Feature extraction

In our system, two types of feature extractions are implemented, the lexicon-based approach and corpus-based approach. The data transformation approach we took is a popular weighting function, Term Frequency–Inverse Document Frequency (TF-IDF) (Manning et al. 2008). Regarding the lexicon-based approach, the features are from the mapping between documents and the selected lexicon. The lexicons we proposed are dictionaries used in NER, and Consumer Health Vocabulary[4](CHV), which connects informal, common words and phrases about health to technical terms used by healthcare professionals. In the corpus-based approach, the features are automatically discovered from the context of documents and they are more specific on certain topics. All words in the documents are the candidates to be the features. We implemented two corpus-based approaches, the bag-of-words model and Rapid Automatic Keyword Extraction (RAKE) (Rose et al. 2010). The bag of words model is a simple information retrieval method and usually used in natural language processing and text classification. The bag of words method uses the occurrence of each word in the corpus as a feature for training a classifier. RAKE is an algorithm to extract features based on some constraints such as the length of a word, the length of a phrase and the occurrence frequency. This study developed a typical feature extraction including three main components, that are "candidate selection" extracting all possible words, phrases, terms or concepts to be feature candidates, "properties calculation" calculating the properties of candidates to indicate the possibility of being a feature, and "scoring and selecting" accumulating the values of the properties of each candidate as a score and identifying the feature set by checking if their score passes the predefined threshold.

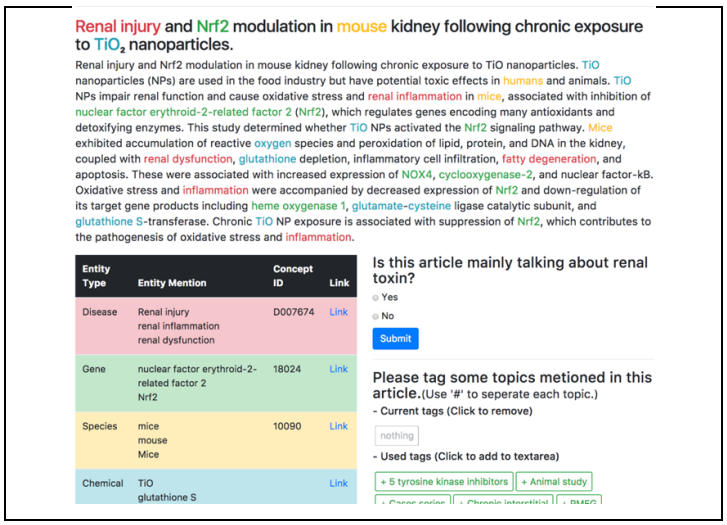

Figure 2. The detail page of ground truth curation

Ground truth curation

Researchers can define at least two classes in our platform, and then provide the ground truth for the further categorization according to which definition of classes the document is most with. To assist the annotator to efficiently curate, a friendly curation page is provided. Firstly, we designed a summary page for the retrieved results and it simply displays some general information of the retrieved articles, like "title", "journal", "year", and "PubMed link". Secondly, we designed a detail page shown in Figure 2. The detail page contains the detailed information of the article under the curation, such as "title", "the content of the abstract", and the corresponding entities. We used a different color to tag different types of entities in the abstract. Red is designed for highlighting entities of diseases, yellow is used for species entities, and blue is for chemical entities. Along with the content of the abstract, we also provided an entity table under the abstract to organize and show the detail of types of entities. By

[4] http://consumerhealthvocab.org/

highlighting different types of entities of the article and providing the summary of entities aside, the annotator can get the idea of the article quickly and easily identify whether the article is their target or not.

Association extraction

Extracting relations between entities can not only help users to understand the corpus easily but also extract the potential associations. Some of the research aimed to detect the potential interactions such as protein-protein or gene-disease interactions. However, this kind of research usually needs to define some rules or to adopt automatic algorithms that are difficult to explain to the people who are not in the machine learning field. Therefore, an association rule-based method is conducted in our system. Agrawal and Srikant (1994) proposed an algorithm to find associate relations efficiently, which is called Apriori algorithm. Apriori algorithm used the breadth-first search and hash tree structure to calculate the frequency of the item set. There are two output results, *"Support"* and *"Confidence"*. Support means the co-occurrence between two item sets, and Confidence means the conditional probability between two item sets. In our system, we attempted to use a rapid way to extract and explain the associations. We provided a blacklist function to allow researchers to label entities with the importance level in their project so that we could adopt the importance level into the ranking of the association result. In the end, our system can provide a sorted list based on the user preference and the co-occurrence of the item set. Researchers can get meaningful information efficiently. After getting some meaningful association by applying Apriori algorithm, our system will visualize the association result by plotting a network graph. In a network graph, a node represents an entity and an edge represents the confidence between two entities.

Socialization

When a project comes to the end, the author of this project can share the outcome of the project with the public including the curation result, categorization result, or association result. The diagram of socialization concept is shown in Figure 3. The categorization result and association result of the public projects can be reused and extended by other teams who have the similar purpose of the projects. This system will be a platform where researchers can share their precious dataset and fully focused on their research goals instead of spending time and labor on the datasets.

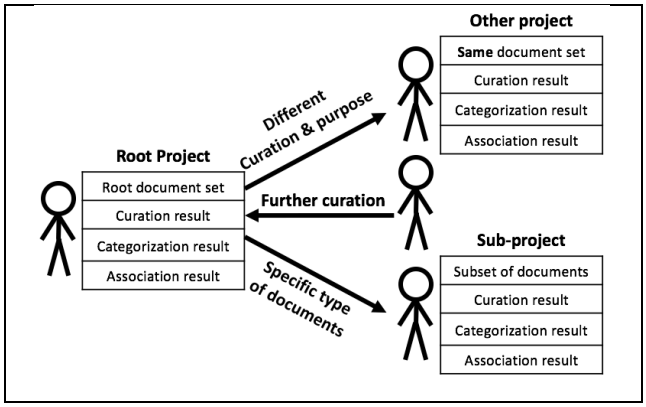

Figure 3. Socialization concept

Case Study – Chronic Kidney Disease

Chronic kidney disease (CKD) is a global issue (Mehta et al. 2015). In Taiwan, the prevalence of CKD is around 9.8 – 11.9% (Kuo et al. 2007). Although CKD patients with old age took up 7.7% of the entire Taiwan population, they had consumed around 15% of the annual National Health Insurance budget (Hwang et al. 2010). The most common cause of CKD is diabetic mellitus, chronic glomerulopathy, hypertension, and chronic interstitial nephritis. Although there is no documented meta-analysis of the actual cause of CKD, it is likely that CKD secondary to nephrotoxic compound exposure took up a

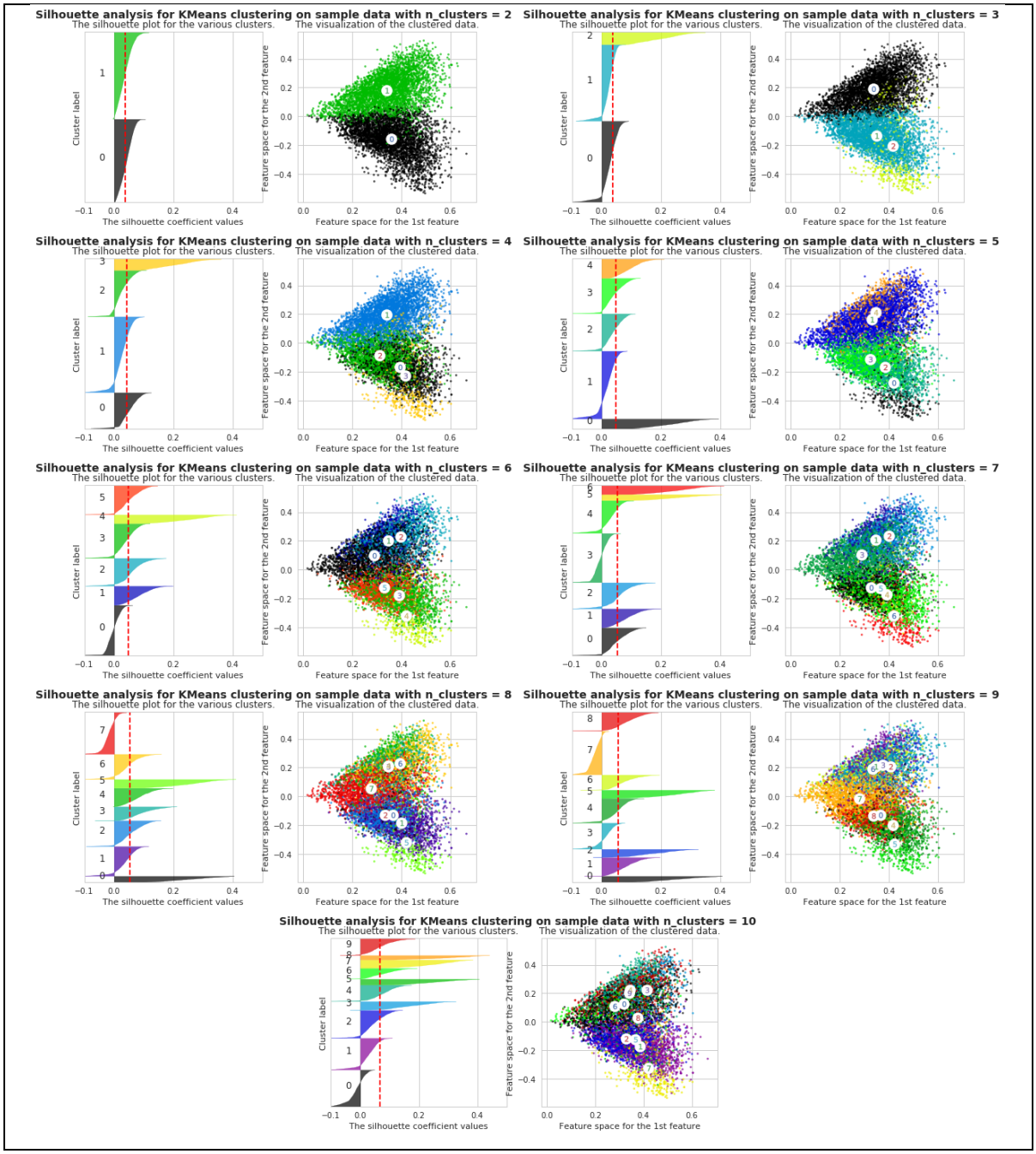

Figure 4. Silhouette analysis and scatter plot of the clustering result

portion of these cases. The process of nephrotoxicity kidney is complicated. The loss of kidney function (renal failure) is the final presentation of the injury of a kidney that due to pre-renal conditions (i.e. hypotension), injuries of the renal vascular system, glomerulopathy, injuries of interstitial/tubular, or obstructive nephropathy. To understand the mechanism of nephrotoxic compounds contributing to renal failure, accurately classified the nephrotoxic agents is crucial.

A research team is formed with a group of medical domain experts who are currently board-certified pathologist and specialized in general pathology, uropathology, gynaecological pathology, and renal pathology. They attempted to create and organize a database of the nephrotoxic related compound that generates renal toxicity. They planned to create a database that includes the reported nephrotoxic agents published in PubMed so that the database can help them to understand the nephrotoxicity of a compound, and facilitate the study of a nephrotoxic agent.

The research team started the project and named it as "Nephrotoxicity project". They defined the purpose of this case is to explore the relationship between compounds of nephrotoxicity that generates renal toxicity. They then imported 30,812 MEDLINE abstract documents collected by using 111 kidney-disease-related keywords such as Azotemia, Glomerulonephritis, and Acute Tubular Necrosis (ATN) from PubMed and export the retrieved results with MEDLINE format. In this project, the experts selected chemical and disease entities from our dictionaries.

Document categorization

After creating the project, our system automatically did K-means clustering and separated this dataset from 2 to 10 clusters. To assess the result, we computed the silhouette coefficient value (Rousseeuw 1987) and extracted two principal components to visualize the scatter plot shown in Figure 4. The mean of silhouette coefficient value for each result is under 0.1 which means that documents are similar to documents from different clusters. Besides, the research team was also hard to distinguish the difference between the 2-clustering result and 3-clustering result. The results of K-means clustering algorithm showed that the dataset is a highly-relevant and specific-oriented corpus, which is no clear way to automatically group.

Since the retrieved documents are all highly relevant to kidney diseases, the various causes of CKD make their retrieved abstracts not only containing nephrotoxicity causing by compounds but also other causing renal toxicity reasons such as genetics issues. To make the dataset only for their purpose, the research team has to firstly classify abstracts with two pre-defined categories "Curated" and "Rejected". The "Curated" category represents that the compound entities, chemicals and herbs, in the documents are related to kidney injury or used to treat kidney diseases. The "Rejected" category represents that the documents which are kidney related but not about kidney diseases or are kidney disease related but not about toxicology. To make the classification work, a sufficient ground truth has to be provided through the interface (Figure 2). In total, professionals annotated 1511 documents as "Rejected" and 361 documents as "Curated" during three months.

While they were providing annotations, they could try different classifiers and feature extraction methods to predict and check whether the size of ground truth is enough to provide them a reliable classification through our platform. For example, Figure 5 shows that when they curated 500 documents, they used the bag of words method to extract features and the random forest as their classifiers. The result of the overall performance at the moment was shown in a figure which contains x-axis representing a different sample size of training data, y-axis representing the score of predicting performance and different color lines to show precision, recall, and f1-score of each category or overall. The different figure represents a different classifier's performance, and they can compare these classifiers' performance. When the number of curated documents increases to 1,000, researchers can fit the predicted model again and observe the improvement shown in Figure 5. Table 1., Table 2., and Table 3 show that the combination of different feature types and learning algorithms. In this case, there is no significant difference between these three feature extraction methods, namely bag of words, RAKE, and CHV. Comparing the learning algorithms, the SVM and the random forest have similar performance and are better than the MNB. The imbalance condition on the categories makes classifiers hard to predict the "Curated" category, especially the MNB. However, all classifiers have good performance on predicting the "Rejected" category which means that the unrelated documents can be effectively removed.

Association extraction

When the classification performance came to an acceptable result by the research team, they selected a feasible classifier to predict the rest of the unannotated documents. The system would allow them to export the predict results in an excel format so that the research team could check the annotation results fast. After that, they used the association extraction function to explore possible relationship between compounds of nephrotoxicity. The association result would be present in a network plot (Figure 6).

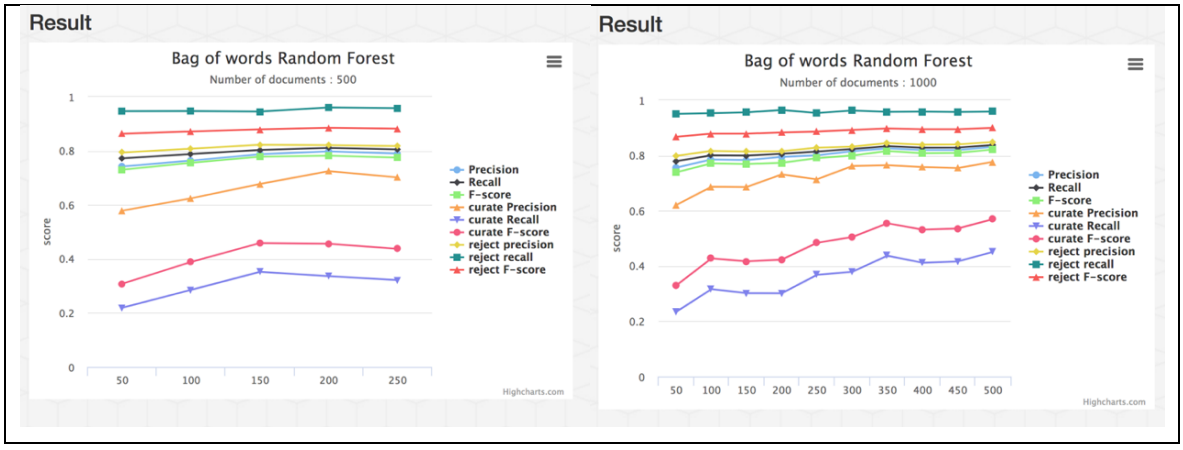

Figure 5. The classification performance with 500 and 700 curated documents

Table 1. Bag of words Result

	Curate precision	Curate recall	Curate F-score	Reject precision	Reject recall	Reject F-score	precision	recall	F-score
MNB	92.86%	7.2%	13.37%	77.41%	99.83%	87.2%	81.1%	77.7%	69.56%
SVM	64.05%	54.29%	58.77%	86.31%	90.43%	88.32%	80.99%	81.8%	81.26%
RF	67.75%	51.8%	58.71%	85.91%	92.26%	88.97%	81.57%	82.59%	81.74%

Table 2. RAKE Result

	Curate precision	Curate recall	Curate F-score	Reject precision	Reject recall	Reject F-score	precision	recall	F-score
MNB	80.49%	9.14%	16.42%	77.69%	99.3%	87.18%	78.36%	77.76%	70.27%
SVM	73.41%	51.25%	60.36%	86.02%	94.17%	89.91%	83.01%	83.92%	82.85%
RF	70.95%	41.27%	52.19%	83.7%	94.7%	88.86%	80.66%	81.93%	80.1%

Table 3. CHV Result

	Curate precision	Curate recall	Curate F-score	Reject precision	Reject recall	Reject F-score	precision	recall	F-score
MNB	91.3%	5.82%	10.94%	77.15%	99.83%	87.04%	80.53%	77.37%	68.85%
SVM	70.69%	45.43%	55.31%	84.6%	94.09%	89.09%	81.27%	82.46%	81.02%
RF	71.78%	40.17%	51.51%	83.5%	95.04%	88.9%	80.7%	81.93%	79.97%

Since researchers' interest is "drug that causes kidney damage or injury", some of the highlighted disease or condition, such as "malignancy remain ongoing problem", "septic renal failure" and "biliary excretion and enterohepatic circulation" are not related with our topic. Besides, some of the compounds are endogenous products, i.e. biochemical compounds that can be found in human body under normal condition. These include "s-creatinine", "nitric oxide", "sodium", "glutathione", etc. The frequent association of these compounds with nephrotoxic is not surprised, since these chemical compounds are an indicator of the human body to kidney injury. Thus, they are not by definition a "toxin". Although some relations which contain these entities are not interesting in researchers, researchers still can quickly find some high frequent relations containing disease entities such as 'Kidney diseases', 'acute renal failures', and 'ischemic acute tubular necrosis' and chemical entities which are not endogenous products such as 'cyclosporin a' and 'cis-diamminedichloro-platinum ii'. Then, researchers can click the relations to fetch all documents which contain this relation. While the research team identifies the associations, their identification could be another run of curation for the system to improve the association extraction. Finally, the research team not only achieve their goal to find associations

between a large and unstructured dataset. While step-by-step processing, they also generated several valuable resources including a dataset of compound-related kidney diseases and the relations between these compounds. In the future, they can share these datasets and results with other professionals through our platform.

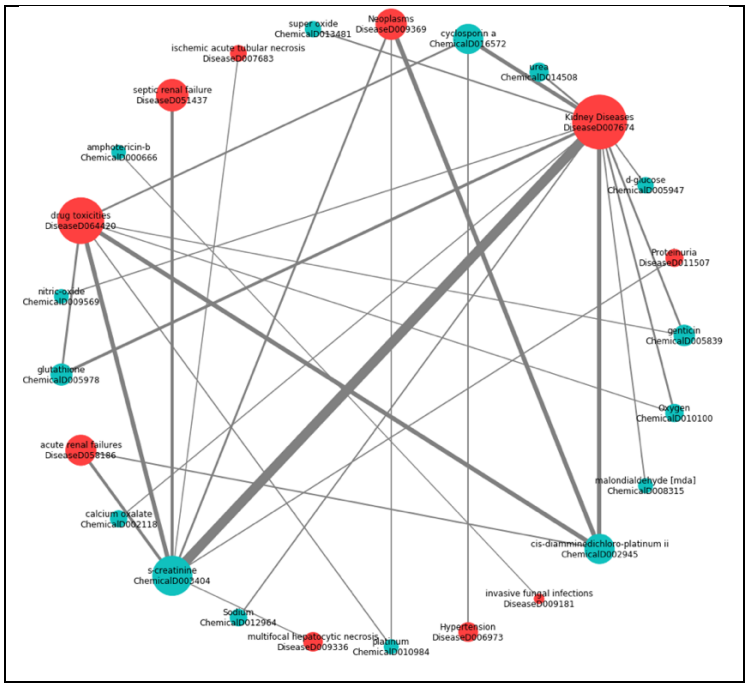

Figure 6. Association network plot

Conclusion

In this study, we presented an integrated web-based text mining platform for analyzing MEDLINE abstracts and helping medical researchers easily manage their project. Name entity recognition helps researchers quickly grasp entities which appear in a document. Document categorization distinguishes different types of documents and researchers can easily focus on a specific type of documents to do further analysis like association extraction. Association extraction effectively explores possible relationships between entities. Besides, the most important feature of this system is that all projects are reusable for other studies by sharing. Medical professionals not only can easily utilize the data mining techniques but also save their time and labor through the dataset sharing mechanism if there were a similar project.

To evaluate this platform, we cooperated with a group of medical professionals and conducted a renal toxicity case to explore the relationship between chemical compounds and renal toxicity. First, this platform automatically did K-means clustering but the results were not helpful for the researchers to distinguish the difference between different clusters. Next, experts used the detail page, went through the relations between these entities and rapidly curated the documents to be the ground truth for the categorization. While annotating documents, experts could also fit some models and track the prediction performance. The prediction performance would improve along with the increase of the number of annotated documents. This platform would show some figures for experts to observe the improvement of different ensemble solutions of features and classifiers. Once the prediction performance reaches an acceptable level which is decided by the experts, experts could use the model to predict the rest of the unannotated documents and then focus on the category they are interested in. In the case study, the best predicting model (RAKE feature type and SVM learning algorithm) had 83.01% precision and 83.92% recall in 10-fold cross-validation. Although the recall of the "Curated" category was not very well (51.25%), the precision was 73.41%, the performance has shown that the prediction of the "Curated"

category had enough confidence to be trusted. Besides, the performance on predicting the "Rejected" category was better with the 86.02% precision and 94.17% recall, which indicates that the classifier could effectively remove the unrelated documents. Finally, the experts used the classifier with the best performance to predict the rest of the unannotated documents and further explored the entity relations in the "Curated" documents. The relationship exploration result are mainly high-frequency relationships, that are mostly about endogenous products from human body to kidney injury. Although most of the extracted relationships are not surprising to researchers in the renal domain, professionals still found some interesting relationships that are not part of endogenous products.

Our system still has some limitations. First, the prediction on the 'Curated' type is weaker than the 'Rejected' type. We think the main reason is that these two categories corpora are imbalanced. Although we have tried some resampling methods, the recall of predicting the 'Curated' type is still not as good as the 'Rejected' type. Under this circumstance, medical researchers need to curate more ground truth corpora to make the predicting performance better. Second, the association relationship extraction algorithm in this system is based on co-occurrence frequency of different entities. This method can effectively find the high frequent relations. However, once the medical researchers wants to find some novel and meaningful relations, this algorithm is hard to detect these relations.

Acknowledgements

This work was sponsored by National Sun Yat-sen University and Kaohsiung Medical University Project in Taiwan: NSYSUKMU 105-I005, and MOST 106-2410-H-004-081 to the first author.

References

Agrawal, R., and Srikant, R. 1994. "Fast Algorithms for Mining Association Rules," in *Proc. 20th Int. Conf. Very Large Data Bases, VLDB, (1215)*, pp. 487–499.

Björkelund, A., Bohnet, B., Hafdell, L., and Nugues, P. 2010. "A High-Performance Syntactic and Semantic Dependency Parser.," in *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations. Association for Computational Linguistics*, pp. 33–36.

Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. "Latent Dirichlet Allocation.," *Journal of Machine Learning Research 3*, pp. 993–1022.

Bravo, À., Piñero, J., Queralt-Rosinach, N., Rautschka, M., and Furlong, L. I. 2015. "Extraction of Relations between Genes and Diseases from Text and Large-Scale Data Analysis: Implications for Translational Research," *BMC Bioinformatics* (16:1), p. 55.

Breiman, L. 2001. "Random Forests," *Machine Learning* (45:1), pp. 1–33.

Cooper, J. W., and Kershenbaum, A. 2005. "Discovery of Protein-Protein Interactions Using a Combination of Linguistic, Statistical and Graphical Information.," *BMC Bioinformatics* (6:1), p. 143.

Cortes, C., and Vapnik, V. 1995. "Support-Vector Networks," *Machine Learning* (20:3), pp. 273–297.

Coulet, A., Shah, N. H., Garten, Y., Musen, M., and Altman, R. B. 2010. "Using Text to Build Semantic Networks for Pharmacogenomics," *Journal of Biomedical Informatics* (43:6), pp. 1009–1019.

Cuingnet, R., Rosso, C., Chupin, M., Lehéricy, S., Dormont, D., Benali, H., Samson, Y., and Colliot, O. 2011. "Spatial Regularization of SVM for the Detection of Diffusion Alterations Associated with Stroke Outcome," *Medical Image Analysis* (15:5), pp. 729–737.

Gerner, M., Sarafraz, F., Bergman, C. M., and Nenadic, G. 2012. "BioContext: An Integrated Text Mining System for Large-Scale Extraction and Contextualization of Biomolecular Events," *Bioinformatics* (28:16), pp. 2154–2161.

He, L., Yang, Z., Lin, H., and Li, Y. 2014. "Drug Name Recognition in Biomedical Texts: A Machine-Learning-Based Method," *Drug Discovery Today* (19:5), pp. 610–617.

Huang, A. 2008. "Similarity Measures for Text Document Clustering," in *Proceedings of the Sixth New Zealand*, pp. 49–56.

Hwang, S. J., Tsai, J. C., and Chen, H. C. 2010. "Epidemiology, Impact and Preventive Care of Chronic Kidney Disease in Taiwan," *Nephrology (Carlton)* (15 Suppl 2:June), pp. 3–9.

Kazama, J. 2001. "Tuning Support Vector Machines for Biomedical Named Entity Recognition," in *Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain-Volume 3. Association for Computational Linguistics*, pp. 1–8.

Kibriya, A. M., Frank, E., Pfahringer, B., and Holmes, G. 2004. "Multinomial Naive Bayes for Text Categorization Revisited," in *Advances in Artificial Intelligence*, pp. 488–499.

Kuo, H. W., Tsai, S. S., Tiao, M. M., and Yang, C. Y. 2007. "Epidemiological Features of CKD in Taiwan," *American Journal of Kidney Diseases* (49:1), pp. 46–55.

Lin, Y. L., Chung, C. Y., Kuo, C. W., and Chang, T. M. 2016. "Modeling Health Hare Q&A Questions with Ensemble Classification Approaches.," in *Proceedings of AMCIS 2016*, pp. 1–10.

Liu, Y., Liang, Y., and Wishart, D. 2015. "PolySearch2: A Significantly Improved Text-Mining System for Discovering Associations between Human Diseases, Genes, Drugs, Metabolites, Toxins and More," *Nucleic Acids Research* (43:W1), pp. 535–542.

MacQueen, J. 1967. "Some Methods for Classification and Analysis of Multivariate Observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297.

Manning, C. D., Raghavan, P., and Schütze, H. 2008. "Scoring, Term Weighting, and the Vector Space Model," in *Introduction to Information Retrieval*, p. 100.

Mccallum, A., and Nigam, K. 1997. "A Comparison of Event Models for Naive Bayes Text Classification," in *AAAI-98 Workshop on Learning for Text Categorization* (Vol. 752), pp. 41–48.

Mehta, R. L., Cerdá, J., Burdmann, E. A., Tonelli, M., García-García, G., Jha, V., Susantitaphong, P., Rocco, M., Vanholder, R., Sever, M. S., Cruz, D., Jaber, B., Lameire, N. H., Lombardi, R., Lewington, A., Feehally, J., Finkelstein, F., Levin, N., Pannu, N., Thomas, B., Aronoff-Spencer, E., and Remuzzi, G. 2015. "International Society of Nephrology's 0by25 Initiative for Acute Kidney Injury (Zero Preventable Deaths by 2025): A Human Rights Case for Nephrology," *The Lancet* (385:9987), pp. 2616–2643.

Neelakantan, A., and Collins, M. 2014. "Learning Dictionaries for Named Entity Recognition Using Minimal Supervision," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 452–461.

Papanikolaou, N., Pavlopoulos, G. A., Theodosiou, T., and Iliopoulos, I. 2015. "Protein-Protein Interaction Predictions Using Text Mining Methods," *Methods* (74:October), pp. 47–53.

Pletscher-Frankild, S., Pallejà, A., Tsafou, K., Binder, J. X., and Jensen, L. J. 2015. "DISEASES: Text Mining and Data Integration of Disease-Gene Associations," *Methods* (74), pp. 83–89.

Rennie, J. D. M., Shih, L., Teevan, J., and Karger, D. R. 2003. "Tackling the Poor Assumptions of Naive Bayes Text Classifiers," in *Proceedings of the Twentieth International Conference on Machine Learning (ICML)-2003)*, pp. 616–623.

Rish, I. 2001. "An Empirical Study of the Naive Bayes Classifier," in *International Joint Conference on Artificial Intelligence*, pp. 41–46.

Rose, S., Engel, D., Cramer, N., and Cowley, W. 2010. "Automatic Keyword Extraction from Individual Documents.," *Text Mining: Applications and Theory*, pp. 1–20.

Rousseeuw, P. J. 1987. "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis.," *Journal of Computational and Applied Mathematics 20*, pp. 53–65.

Rzhetsky, A., Iossifov, I., Koike, T., Krauthammer, M., Kra, P., Morris, M., Yu, H., Dubou, A., Wilbur, W. J., Hatzivassiloglou, V., and Friedman, C. 2004. "GeneWays : A System for Extracting , Analyzing , Visualizing , and Integrating Molecular Pathway Data," *Journal of Biomedical Informatics* (37:1), pp. 43–53.

Stenetorp, P., Pyysalo, S., and Topi, G. 2012. "BRAT: A Web-Based Tool for NLP-Assisted Text Nnnotation," in *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics*, pp. 102–107.

Stephens, M. J., Palakal, M. J., Mukhopadhyay, S., Raje, R. R., and Mostafa, J. 2001. "Detecting Gene Relations from Medline Abstracts," in *Pacific Symposium on Biocomputing*, pp. 483–496.

Tong, S., and Koller, D. 2001. "Support Vector Machine Active Learning with Applications to Text Classification," *Journal of Machine Learning Research* (2:Nov), pp. 45–66.

Wei, C., Kao, H., and Lu, Z. 2013. "PubTator: A Web-Based Text Mining Tool for Assisting Biocuration," *Nucleic Acids Research* (41(W1)), pp. 518–522.

Wu, Q., Ye, Y., Zhang, H., Ng, M. K., and Ho, S.-S. 2014. "ForesTexter: An Efficient Random Forest Algorithm for Imbalanced Text Categorization.," *Knowledge-Based Systems* (67), pp. 105–116.

Xie, P., and Xing, E. P. 2013. "Integrating Document Clustering and Topic Modeling," in *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence*, pp. 694–703.

Zhang, W., Yoshida, T., and Tang, X. 2008. "Text Classification Based on Multi-Word with Support Vector Machine," *Knowledge-Based Systems* (21:8), pp. 879–886.

Zhu, F., Patumcharoenpol, P., Zhang, C., Yang, Y., Chan, J., Meechai, A., Vongsangnak, W., and Shen, B. 2013. "Biomedical Text Mining and Its Applications in Cancer Research," *Journal of Biomedical Informatics* (46:2), pp. 200–211.