A hybrid kernel based method for relation extraction and gene-disease interaction network construction

Lei Hua[1] , Changqin Quan[2], Fuji Ren[3]

[1]School of Computer and Information, HeFei University of Technology, China
[2]Graduate School of System Informatics, Kobe University, Japan
[3]The University of Tokushima, 770-8506, Japan
hualeilxf@163.com, quanchqin@hfut.edu.cn, ren@is.tokushima-u.ac.jp

Abstract: The existence of a large number of information in biological literatures motivated the using of data mining methods to extract biomedical relations automatically. Especially the kernel based machine learning methods have been widely used to extract protein-protein interactions (PPI) in the field of biomedical nature language processing (BioNLP). In this paper, we propose a method that combines kernels based on edit-distance and cosine similarity to deal with biomedical relation extraction (RE) task, and then we evaluate our method on AIMED and BioInfer corpus through modifying the kernel of LibSvm. The proposed method is superior to the existing relation extraction systems on BioInfer corpus, and the results show that the proposed kernel obtains higher precision than the existing methods, which illustrates its capacity of learning. On the other hand, we observe that the f-scores are improved when multiple corpora are used for training. Moreover, we apply the trained RE system on publicly available corpus which crawled from PubMed database with the keyword “breast cancer”, through the extracted gene-disease and gene-gene interactions, we construct the gene interaction network, then a network analysis method which combines degree centrality and betweenness centrality is applied to extract breast cancer disease related gene.

Keywords: Relation extraction; Edit-distance kernel; gene-disease interaction network.

1 Introduction

Biomedical relations play an important role in biologic processes and are widely researched in the field of BioNLP. PPI and gene disease interactions have captured much interest among the study of biomedical relations especially, because relation between biomedical entities can help understand the genetics in human health and provide useful information.

In general, RE systems can divide into two aspects: extracting relations based on structure databases and based on text-mining methods. Most of the structure databases are constructed manually, like BIND [1], MINT [2], IntAct [3], and Swiss–Prot [4]. However it is difficult to maintain an interaction database with the rapid growth of biological literatures. Therefore, the implementation of text-mining methods become an important research area. Text-mining methods focus on the features of word information (such as part-of-speech, stem) and the dependency relations or structure features in sentences, and the usage of nature language processing (NLP) tools can make a sentence more explicit, the problem is how the algorithms take advantage of the information generated from NLP tools, and thus, some kernel methods are proposed. Kernels based methods have been extensively researched in various RE fields (like newspaper text and biomedical) and achieved better result than pattern match [5, 6] or rule based methods [7]. Kernel methods can be used to calculate the similarity of two input vectors. Through modifying the kernel of existing methods such as SVM or LibSvm, we can easily perform classification in high-dimensional space. However the proposed kernel based methods including bag-of-words kernel [8], all-paths kernel [9], subset tree kernel [10], edit-distance kernel [11], walk-weighted subsequence kernel [12], graph kernel [13] etc. can partly calculate the similarity of two input text, consequently, it is necessary to combine the kernels [14] to gather more available information.

Gene interaction network has been applied in many biologic processes recently, for example, the interaction network can help find functional modules [15] among gene entities, and the network can also find the functionally similar genes with some initial seed gene entities [16, 17]. The proposed network analysis is different from previous approaches mainly in one aspect, we focus on the most relevant gene entities to breast cancer through ranking the extracted gene entities. And thus, we care more about the centrality of a gene in an interaction network, and the nodes at the border of the network are ignored, which has been demonstrated that can help improving the accuracy.

In this paper, we introduced a hybrid kernel based relation extraction approach and an interaction network analysis method. Different from the existing kernel methods, we combine kernels based on edit distance and cosine similarity, and implement our method through modifying the kernel of LibSvm [18]. Moreover, we apply the trained RE system on PubMed corpus, a gene-disease interaction network can be constructed. We also propose a method that can effectively measure the importance of a gene in an interaction network to extracted breast cancer related gene entities.

In the evaluation, we apply our method on AIMED and BioInfer standard corpus and compare our results with

other RE systems, the experimental results illustrate that the proposed kernel method achieved better f-scores than other RE systems on BioInfer corpus, and obtained the highest accuracy by comparing with the existing published systems.

2 Materials and Methods

2.1 Dependency parsing for similarity measure and relation extraction

Dependency parsers [7, 10, 11] are widely used in RE task, and a parser can take a sentence as a sequence of words as input, and output the syntactic structures and semantic relationships among words. To extract relation between two entities, the main hypothesis is that the shortest path [11, 19] between two gene entities in parser tree contained useful information for RE.

Figure 1 demonstrates the dependency parsing tree of sentence *"In vitro and in nonproliferating cells, p27 associates with and inhibits cyclin/cycin-dependent kinase holoenzymes containing either CDK4, CDK6, or CDK2."* Different from the previous shortest path kernel, in our method, we combine the words and the dependency relations (for example dependency relations such as *"dep"*, *"prep"* are included in shortest path) which showed to be more effective.

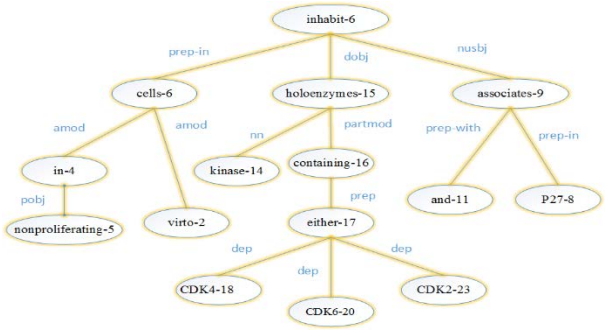

Figure 1. The dependency tree of the sentence "*In vitro and in nonproliferating cells, p27 associates with and inhibits cyclin/cycin-dependent kinase holoenzymes containing either CDK4, CDK6, or CDK2.*"

From the perspective of machine learning method, relation extraction problem can be considered as a binary classification problem. For example, if there are *n* gene entities in a sentence, we will have $n * (n - 1)/2$ pairs of gene entities. Considering the example in Figure 1, there are four entities (p27, CDK4, CDK6 and CDK2) in the example sentence, so we will have six pairs of entities, in general, there is only one path between two entities, however, there are more than one paths between entities sometimes, and we choose the shortest path in this case. The shortest path of entities "CDK2" and "p27" is extracted as follow:

CDK2-dep-either-prep-containing-partmod-holoenzymes-dobj-inhabit-nusbj-assocaiates-prep-in-p27

To better measure the similarity between two input shortest paths and reduce the sparseness of data, we replaced all the gene pairs as special symbol "*GENE*". Therefore the example path can be replaced as:

GENE-dep-either-prep-containing-partmod-holoenzymes-dobj-inhabit-nusbj-assocaiates-prep-in-GENE

2.2 Kernel

2.2.1 Cosine similarity and edit-distance kernels

As mentioned before, in order to implement the classification algorithm to extract relations, an appropriate kernel function should be defined to measure the similarity of two input objects. In this paper, we use the shortest path to generate kernels, however, the different usages of the shortest path can produce different kernels, and we utilize the shortest path to generate cosine similarity and edit-distance kernels in this work. The extracted shortest paths in our method are different from the published shortest path based methods, because we include the dependency relations in our shortest path while most of other shortest path kernels only keep the words between entities [7, 14], as dependency relations play an important role in RE task, the extracted shortest path can not only capture the surface words information but also can obtain the sentence structure and semantic information.

2.2.2 Definition of mathematic symbols for kernels

Suppose W represents all of the words (the dependency relations like *"dep"*, *"prep"* are all considered as single words for convenient) in our word dictionary, N is the sum of words, and each words in dictionary will be assigned with an index from 1 to N for identification. Consequently, the extracted shortest path can be represented as a vector by mapping each words with an index in the dictionary. Specifically p and q are the two input shortest path vectors, $lenp$ and $lenq$ are the lengths of p and q.

2.2.3 Cosine similarity kernel

The cosine similarity described here is an extension of bag-of-word model, the cosine similarity can calculate the common parts of two input vectors and mainly be used to get terms information. The cosine similarity can be defined in equation (1) where f is an indicator function in equation (2).

$$sim\_cos = \frac{1}{lenq * lenp} \sum_{i=0}^{lenp-1} \sum_{j=0}^{lenq-1} f\{p[i], q[j]\} \quad (1)$$

$$f\{p[i], q[j]\} = \begin{cases} 1, & p[i] = q[j] \\ 0, & p[i] \neq q[j] \end{cases} \quad (2)$$

2.2.4 Edit-distance kernel

Cosine similarity ignores the order of words and only takes advantage the common words. Different from cosine similarity, edit-distance can help capture the structure information of two input vectors.

Edit-distance (also called Levenshtein distance) can be simply calculated with a dynamic programming.

Suppose $dis[i][j]$ represents the edit-distance of vectors $p(0:i)$ and $q(0:j)$, where $p(0:i)$ represents the sub-vector of p form index 0 to i. Before running the dynamic programming, we should initial the boundary value of $dis[i][j]$, which are showed in equation (3)(4). After the initial step, the $dis[i][j]$ can be calculated recursively by equation (5).

$$dis[i][0] = i \ \ (i = 0,1, \dots lenp) \tag{3}$$

$$dis[0][j] = j \ \ (j = 0,1, \dots lenq) \tag{4}$$

$$dis[\mathrm{i}][\mathrm{j}] = \begin{cases} \min \begin{pmatrix} dis[i][j-1], \\ dis[i-1][j], \\ dis[i-1][j-1] \end{pmatrix} + 1 \\ if \ \ f\{p[i], q[j]\} \neq 1 \\ \mathrm{dis}[i-1][\mathrm{j}-1] \\ if \ \ f\{p[i], q[j]\} = 1 \end{cases} \tag{5}$$

To better combine the different kernels, edit-distance should be normalized through equation (6). $\lambda 1$ is a hyper parameter which can be modified during the experiments, and experiments result show that it is better to choose $\lambda 1$ less than 1.

$$sim\_edit(p,q) = e^{-\lambda 1 * dis[lenp][lenq]} \tag{6}$$

2.2.5 Hybrid kernel

We propose the hybrid kernel based on the motivation that combination of kernels can make use of all available information, the hybrid kernel is defined by equation (7). $\lambda 2$ is a multiplicative of cosine kernel, this parameter can help to quantify the information which obtained from cosine kernel to final hybrid kernel.

$$hybrid = sim\_edit(p,q) + \lambda 2 * sim\_cos(p,q) \tag{7}$$

2.3 Network analysis for extracting diseased-related gene entities

Based on the extracted gene-gene and gene-disease relations, an interaction network can be constructed. A network can be represented as a direct graph, each node in this graph is a gene entity. Centrality is used to measure the importance of a node in the network in this paper.

2.3.1 Degree centrality

Degree centrality [20] can measure the central tendency of a node in a network (graph). A node will be more important and contains more energy if more edges are connected to this node. Suppose a graph can be represented as an adjacency matrix M, where $M_{ij} = 1$ if there is a relation between gene entity i and j, and the degree centrality $\mathrm{C}_D(i)$ of node i can be calculated in equation (8)

$$C_D(i) = \sum_{j=1} M_{ij} \tag{8}$$

2.3.2 Betweenness centrality

Betweenness centrality [21] reflects the ability of a node to control other nodes, in other words, it measures the ability of a node to take charge of the resources in a network, the more times a specific node on the shortest paths of other pair of nodes, the more resources this node will control, the betweenness centrality $C_B(i)$ of node i can be describe in equation (9), where δ_{st} is the sum of the shortest paths from s to t, and the $\delta_{st}(i)$ is the sum of shortest paths from s to t passing through i.

$$C_B(i) = \sum_{s \neq t \neq i} \frac{\delta_{st}(i)}{\delta_{st}} \tag{9}$$

2.3.3 Normalized centrality

To better measure the importance of a node in the network, we apply normalized centrality method to make use of degree and betweenness centrality. The normalized centrality $C_W(i)$ is showed in equation (10), where D_MAX is the maximum value of degree centrality and B_MAX is the largest value of betweenness centrality.

$$C_W(i) = \frac{C_D(i)}{D\_MAX} + \frac{C_B(i)}{B\_MAX} \tag{10}$$

3 Result and discussion

3.1 Datasets

The datasets include two parts. Dataset 1 is used to evaluate the proposed kernel method, and Dataset 2 is utilized to extract gene-disease relations.

Dataset 1: This dataset contains three parts of standard corpora including AIMED [22], BioInfer [23] and IEPA [24]. AIMED are manually tagged by Bunescu et al. which includes about 200 medical abstracts and is considered as a standard dataset for PPI task. BioInfer is developed by Turku BioNLP group and IEPA dataset is tagged by Ding et al. In the experiments, AIMED and BioInfer are used to test our methods while the IEPA is used to co-training while training LibSvm model. The main reason we choose AMIED and BioInfer as our evaluation corpus is that the scales of AMIED and BioInfer datasets are bigger than other standard corpus like LLL [25] and HPRD [7], which can help to learn more data structure[1]. Moreover, we preprocess the five corpora into XML files for the purpose of research and convenient[2].

Dataset 2: To extract disease related genes/proteins, we crawled 28039 abstracts from PubMed with key word *"breast cancer"*(Apache Axis2 provides us with an easy way to visit PubMed database [3]), the preprocessing corpus include 190171 sentences, an gene name dictionary crawled from OMIM dataset is used to identify the gene name entities in sentences, and all the tidy data, gene name dictionary and our

[1]http://corpora.informatik.hu-berlin.de/

[2]https://github.com/coddinglxf/RelationExtractionMatrerial/tree/master/stav_xml_new

[3]http://www.ncbi.nlm.nih.gov/entrez/eutils/soap/v2.0/DOC/esoap_java_help.html#ex_run_ eGquery

source code can be found at here[4]. The trained RE system on BioInfer will be applied on this dataset to extract gene-disease relations.

3.2 Evaluation of kernel based method on Aimed and BioInfer corpus

3.2.1 Evaluation method

We mainly use precision (p), recall (r) and f-score (f) to evaluate the performance of our kernel method. The f can be defined in equation (11), and 10 cross validation (10-fold CV) method is used to calculate the average f-scores.

$$\mathrm{f} = \frac{2 * p * r}{p + r} \qquad (11)$$

3.2.2 The influence of different values of $\lambda 1$

In general, a well-defined kernel function should be positive definite, while the edit-distance is not always positive definite which had been proved by Cortes [26], however, we can adjust the value of $\lambda 1$ in equation (6) to make our result much better. We utilize 10-fold CV method on AIMED corpus to evaluate the influence of different values of $\lambda 1$ based on edit-distance kernel. Empirical results show that it is better to choose the value of $\lambda 1$ under 2. To get the suitable $\lambda 1$ and $\lambda 2$, we employ the grid search method, we set the $\lambda 1$ and $\lambda 2$ range from [0.5, 2] and [0.00001, 0.1] respectively, the tuned parameters through cross-validation experiments are demonstrated in Table 1.

Table 1. The suitable parameters selected by grid search method based on hybrid kernel

Dataset	$\lambda 1$	$\lambda 2$
AIMED	0.5	0.001
BioInfer	0.5	1e-5
AIMED + IEPA	0.75	0.1
BioInfer + IEPA	0.5	1e-4

3.2.3 The multiple corpora for training

On the evaluations of our hybrid kernel, we find that more labeled data added in train dataset can help improve f-score (Miwa et al. [27] also showed that better results can be obtained with multiple corpora). In this experiment, we choose IEPA as co-training corpus and put it in train dataset when we train LibSvm on AIMED and BioInfer corpus. We utilized the same grid search method, and find the best parameters for co-training experiment which are showed in Table 1.

[4]https://github.com/coddinglxf/RelationExtractionMatrerial/tree/master/LIBSVM_FOR_COOO0O0O00L

3.2.4 Results and discussion for the proposed kernel base method

Results: Table 2 illustrates the f-scores of the proposed kernel based method, which includes three aspects: edit-distance kernel only, hybrid kernel and hybrid kernel with co-training of IEPA corpus. We also compare our methods with other RE systems in Table 3.

Table 2. The f-scores of hybrid methods on AIMED and BioInfer corpus

Dataset	Edit-distance kernel	Hybrid kernel	Hybrid kernel +IEPA
AIMED	59.8	62.5	62.9
BioInfer	70.0	70.3	71.6

Table 3. Compare with other RE methods

	AIMED			BioInfer		
Kernel based methods	p	r	f	p	r	f
Shallow kernel [28]	60.9	57.2	59.0	-	-	-
Graph kernel [9]	52.9	61.8	56.4	-	-	-
TSVM + edit distance [11]	58.4	61.2	59.6	-	-	-
Hybrid kernel [14]	55.0	**68.8**	60.8	65.7	**71.1**	68.1
Multiple features [27]	-	-	**64.2**	-	-	67.6
Walk weighted kernel [12]	61.4	53.3	56.6	61.8	54.2	57.6
Our hybrid kernel only method	**85.3**	49.5	62.5	**86**	60	70.3
Our hybrid kernel + IEPA method	83.1	50.9	62.9	83.7	62.7	**71.6**

Discussion: In this paper, we address the problem of RE based on kernels, especially the hybrid kernel. As we mentioned before, we consider that the cosine kernel provides us with terms information while the edit-distance kernel generates structure information, the combination of them can help to gather these two useful information. From Table 2, we can see that hybrid kernel performed better than single kernel (2.7% and 0.3% improvement of f-scores on AIMED and BioInfer in Table 2), the improved results have confirmed our motivation of using hybrid kernels. On the other hand, empirical results showed that the multiple corpora can help improve the f-scores (0.4% and 1.3% f-scores improvement on AIMED and BioInfer respectively in Table 2). Our opinion is that more labeled train data can help classification algorithms to learn more data structures, so the recall of the test data can improve (1.4 and 2.7% improvement of recall on AIMED and BioInfer corpus in Table 3), while the learned data structures from multiple corpus may not suitable for current test data, and thus the precision of the system is declined (2.2% and 2.3% decline of precision on AIMED and BioInfer corpus in Table 3). The finally

f-scores which is the harmonic of recall and precision showed that with combined information can enhance the result.

When comparing with other kernel based method in Table 3, the proposed hybrid method on BioInfer corpus outperformed the exiting approaches and has 2.2% f-scores improvement in f-scores of Miwa et al. (2009a)'s hybrid kernel. As for AIMED corpus, our proposed approach is superior to any other supervised methods except for Miwa et al. (2009b)'s multiple features and parsers based method, and has 1.9% f-scores improvement of Erkan et al. (2007)'s semi-supervised approach which also based on edit-distance kernel. Moreover, the experiments results showed that our method obtained *considerably higher precision* than the existing approaches. The hybrid method achieved 23.9% and 20.2% improvements when comparing with Kim et al. (2010)'s walk weighted subsequence kernel, Miwa et al. (2009b)'s multiple features and parsers based method on AIMED and BioInfer corpus respectively, which illustrates its capacity of learning precision model.

We sum up the results of our experiments on RE task. (1) Edit distance is a simple but efficient kernel. (2) The combination of kernels is necessary to gather up more information, and in general, a hybrid kernel performs much better than a single kernel. (3) Multiple corpora can help to improve results.

3.3 Results and discussion for gene-disease interaction network analysis

Figure 2 is the outline of interaction network we constructed from the extracted gene-disease or gene-gene relations, there are 720 nodes in the network, and each yellow node in this figure represents a gene entity except for the most central node which represents the breast cancer, and the red lines in figure stand for the relations among gene entities and disease. On the analysis of network, we mainly focus on the centrality of nodes in the network including degree centrality, betweenness centrality and our proposed normalized centrality.

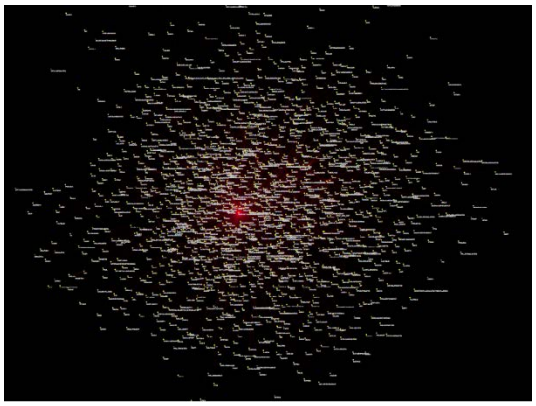

Figure 2. The outline of the derived interaction network of related genes and breast cancer

3.3.1 The result of degree centrality analysis

Degree centrality reflects the centrality tendency of a node in a network. The higher degree centrality of a node has, the more important this node will be. Figure 3 illustrates the outline of the interaction network measured by the degree centrality. The size and the color of the nodes in this figure reflect the values of degree centrality. In general, the deeper the color, the bigger the degree centrality value is. To show the network clearly, we enlarged part of the network, which can be found at the top right corner of Figure 3.

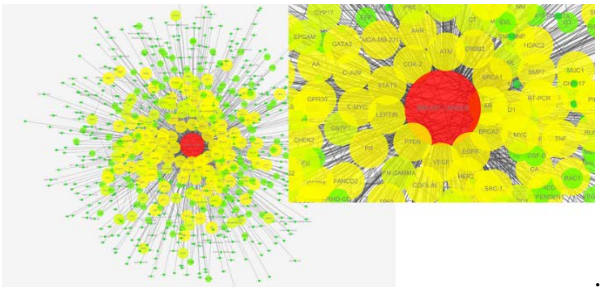

Figure 3. The interaction network of related genes and breast cancer measured by degree centrality

We ranked the extracted gene list according to the degree centrality value of each nodes. The top 10 with highest degree centrality are showed in Table 4, to better describe and understand the extracted genes, we choose HUGO database[5] to normalize the gene names which also can be found at Table 4.

Table 4. The top 10 genes of degree centrality

Gene Name	Degree	Description Of Gene Name
BRCA1	47	breast cancer 1, early onset
EGFR	24	epidermal growth factor receptor
PR	23	progesterone receptor
MDA-MB-231	21	-
D1	21	cyclin d1
ERBB2	18	erb-b2 receptor tyrosine kinase 2
VEGF	18	vascular endothelial growth factor
STAT3	17	signal transducer and activator of transcription 3
HER2	15	human epidermal growth factor receptor-2
EGF	14	epidermal growth factor

[5]http://www.genenames.org/

3.3.2 The result of betweenness centrality analysis

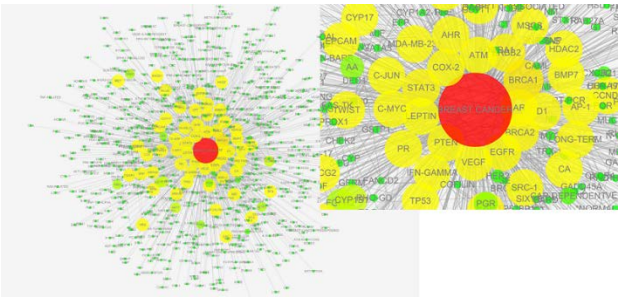

Figure 4. The interaction network of related genes and breast cancer measured by betweenness centrality

Different from the degree centrality, betweenness centrality reflects a node's ability of controlling the other nodes in a network. Figure 4 shows the network measured by betweenness centrality, similar to the analysis of degree centrality, we choose the top 10 gene entities with the highest betweenness centrality which are demonstrated in Table 5.

Table 5. The top 10 genes of betweenness centrality

Gene Name	Betweenness	Description Of Gene Name
BRCA1	0.0220	breast cancer 1, early onset
ERBB2	0.0085	erb-b2 receptor tyrosine kinase 2
STAT3	0.0066	signal transducer and activator of transcription 3
BRCA2	0.0065	breast cancer 2, early onset
D1	0.0054	cyclin d1
BRMS1	0.0051	breast cancer metastasis suppressor 1
EGF	0.0049	epidermal growth factor
EGFR	0.0048	epidermal growth factor receptor
AHR	0.0044	aryl hydrocarbon receptor

3.3.3 The result of normalized centrality analysis

Table 6. The top 15 genes of normalized centrality

Gene Name	Normalized	Description	Relevance
BRCA1	0.1431	breast cancer 1, early onset	YES
ERBB2	0.0552	erb-b2 receptor tyrosine kinase 2	YES
EGFR	0.0491	epidermal growth factor receptor	YES
D1	0.0475	cyclin d1	YES
STAT3	0.0467	signal transducer and activator of transcription 3	YES
BRCA2	0.0411	breast cancer 2, early onset	YES
PR	0.0374	progesterone receptor	YES
MDA-MB-231	0.0367	-	YES
EGF	0.0366	epidermal growth factor	YES
AHR	0.0333	aryl hydrocarbon receptor	YES
VEGF	0.0312	vascular endothelial growth factor	YES
AR	0.0286	androgen receptor	YES
COX-2	0.0273	cytochrome c oxidase subunit	YES
BRMS1	0.0269	breast cancer metastasis suppressor 1	YES
PTEN	0.0266	-	YES

A normalized centrality proposed in equation (10) can gather the information provided by degree and betweenness centrality, Table 6 is the top 15 gene entities through ranking the normalized centrality since we want to find the most related gene entities to breast cancer. We mainly use OMIM database to judge the extracted gene entities in Table 6 whether relating to breast cancer. "BRCA1", "BRCA2", "ERBB2", "AR", "BRMS1" and "PTEN" are the most common breast cancer related genes which can be verified at here[6], the rest of gene entities all can be confirmed by OMIM database.

3.3.4 Discussions for analysis of interaction network

The five gene entities including "BRCA1", "ERBB2", "STAT3", "D1" and "EGF" in Table 5 are corresponded to the degree centrality results in Table 4. The results illustrated that the degree and betweenness centrality are in common to some extent which is the motivation to employ normalized centrality. The top 15 gene entities in Table 6 are almost the combination results of degree and betweenness centrality, which further explains the rationality of the normalized centrality analysis. By ranking the normalized centrality, the system can avoid the mistakes by discarding the entities which at the border of the network. And the results showed that the extracted gene entities are all related to the target disease, thus, so normalized centrality method is robust, and can achieve high accuracy.

4 Conclusion

In this paper, we propose a kernel based method for RE

[6] http://ghr.nlm.nih.gov/condition/breast-cancer

task and employ the method on free text to extract the most relevant gene entities to a specific disease. Comparing with prior work, our kernel based method is easy to implement, and we evaluated our methods on AIMED and BioInfer benchmark datasets.

Experimental results showed that our hybrid kernel outperforms the existing approaches, moreover, our method achieved considerably higher precision, which illustrates its capacity of learning accuracy model.

We also applied our RE system on PubMed datasets to extract the relations among gene entities and breast cancer, and proposed a hybrid centrality method which can measure the importance of a node in a network, the extracted gene entities illustrate that our proposed methods are effective and robust.

Acknowledgements

This research has been partially supported by National Natural Science Foundation of China under Grant No.61472117, No.61203312, and the National High-Tech Research & Development Program of China 863 Program under Grant No.2012AA011103 and the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry, and Key Science and Technology Program of Anhui Province, under Grant No. 1206c0805039.

Reference

[1] Bader GD, Donaldson I, Wolting C (2003) Bind:the biomolecular interaction network database. Nucleic Acids Research 31: 248-250.

[2] Zanzoni A, Montecchi-Palazzi L, Quondam M X (2002) Mint:A molecular interaction database. FEBS Letters 513: 135-140.

[3] Kerrien S, Aranda B, Breuza L (2002) The IntAct molecular interaction database in 2012. Nucleic Acids Research 40: 841-846.

[4] Bairoch A, Apweiler R (2000) The swiss–prot protein sequence database and its supplement trembl in 2000. Nucleic Acids Research 28: 45-48.

[5] Ono T, Hishigaki H, Tanigami A (2001) Automated extraction of information on proteinprotein interactions from the biological literature. Bioinformatics 17: 155-161.

[6] Yakushiji A, Miyao Y, Tateisi Y (2005) Biomedical information extraction with predicate argument structure patterns. Proceedings of the Eleventh Annual Meeting of the Association for Natural Language Processing 93-96

[7] Fundel K, Ku¨ffner R, Zimmer R (2007) RelEx–Relation extraction using dependency parse trees. Bioinformatics 23: 365-371.

[8] Sætre R, Sagae K, Jun'ichi Tsujii (2007) Syntactic features for protein-protein interaction extraction. LBM (Short Papers).

[9] Airola A, Pyysalo S, Bjo¨rne J (2008) All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. BMC bioinformatics 9.

[10] Moschitti A (2006) Making Tree Kernels Practical for Natural Language Learning. EACL 113-120.

[11] Erkan G, Ozgu¨r A, Radev D R (2007) Semi-supervised classification for extracting protein interaction¨ sentences using dependency parsing. EMNLP-CoNLL 7: 228-237.

[12] Kim S, Yoon J, Yang J (2010) Walk-weighted subsequence kernels for protein-protein interaction extraction. BMC bioinformatics 11: 107.

[13] Airola A, Pyysalo S, Bj¨orne J (2008) A graph kernel for protein-protein interaction extraction. Proceedings of the workshop on current trends in biomedical natural language processing. Association for Computational Linguistics 1-9.

[14] Miwa M, Sætre R, Miyao Y (2009a) Proteinprotein interaction extraction by leveraging multiple kernels and parsers. International journal of medical informatics 78: e39-e46.

[15] Spirin V, Mirny L A (2003) Protein complexes and functional modules in molecular networks. Proceedings of the National Academy of Sciences 100: 12123-12128.

[16] Chen J Y, Shen C, Sivachenko A Y (2006) Mining Alzheimer disease relevant proteins from integrated protein interactome data. Pacific Symposium on Biocomputing 6: 367-378.

[17] Ozgu¨r A, Vu T, Erkan G (2008) Identifying gene-disease associations using centrality on a literature¨ mined gene-interaction network. Bioinformatics 24: i277-285.

[18] Chang C C, Lin C J (2011) LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2: 27.

[19] Bunescu R C, Mooney R J (2005) A shortest path dependency kernel for relation extraction. Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. Association for Computational Linguistics 724-731.

[20] Freeman L C (1979) Centrality in social networks conceptual clarification. Social networks, 1979, 1: 215-239.

[21] Freeman L C (1977) A set of measures of centrality based on betweenness. Sociometry 40: 35-41.

[22] Bunescu R, Ge R, Kate R J (2005) Comparative experiments on learning information extractors for proteins and their interactions. Artificial intelligence in medicine 33: 139-155.

[23] Pyysalo S, Ginter F, Heimonen J (2007) BioInfer: a corpus for information extraction in the biomedical domain. BMC bioinformatics 8: 50.

[24] Ding J, Berleant D, Nettleton D (2002) Mining MEDLINE: abstracts, sentences, or phrases. Proceedings of the pacific symposium on biocomputing 7: 326-337.

[25] N´edellec C (2005) Learning language in logic-genic interaction extraction challenge. Proceedings of the 4th Learning Language in Logic Workshop (LLL05) 7: pages 31-37

[26] Cortes C, Haffner P, Mohri M (2004) Rational kernels: Theory and algorithms. The Journal of Machine Learning Research 5: 1035-1062.
[27] Miwa M, Sætre R, Miyao Y (2009b) A rich feature vector for protein-protein interaction extraction from multiple corpora. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1. Association for Computational Linguistics 121-130.
[28] Giuliano C, Lavelli A, Romano L (2006) Exploiting shallow linguistic information for relation extraction from biomedical literature. EACL 18: 401-408.