Research of Extracting Relations and Constructing International Network From News

Jun Wang

A Thesis submitted to the University of Tokushima in
partial fulfillment of the requirements for the degree of
Doctor of Philosophy

March, 2015

Department of Information Science and Intelligent Systems
Graduate School of Advanced Technology and Science
The University of Tokushima

CONTENTS

List of Figures

List of Table

Acknowledgements

I feel very fortunate that I can have the opportunity to study as a doctoral student in the University of Tokushima. First, I would like to thank my advisor Prof. Fuji Ren. With his patient and professional guidance, I spent a nice time at Tokushima that can study interesting topics in natural language processing.

I would also thank Prof. Yixin Zhong at Beijing University of Posts and Telecommunications who gave me lots of suggestions during my studying time.

Thanks to all the members of A1 group and my friends as well. We discuss difficulties about work and share the happiness about any progress. The unity and friendly atmosphere always assist me in studying.

Thanks to my family. They support me consistently during these years of studying; even they are suffering from illness.

Finally, I would say thanks to Prof. Kenji Terada and Prof. Masami Shishibori who review my thesis despite their busy schedule.

Abstract

With the rapid development of communications and information technology, especially the emergence of Internet, information travels faster and easier than ever before. Internet extends the range and scope of communications around the world. Countries become closer and form a huge and complex international network. Based on the background, researchers hope to study international relations through the view of network, analyze the structure of international network, understand effect of different relations on international situation and reveal the evolution of international relations. It is called the research of international network which arouse more and more attention recently. In this thesis, we describe our research of extracting relations and constructing international network, an integral system that can construct international network from news is proposed as well.

One thing should be pointed out that network approach is a formalization method. It has some natural defect such as loss of information while facing with the non-formalization problem of studying international relations. The network approach cannot solve all the problems of international relations as well. Therefore, the main target of our research is not to offer a package that can solve everything. We try to innovate based on current work and provide a better formalization approach to deal with international relations. Our research can offer a tool to assist researchers of international relations. On the other hand, it provides an approach to help common users getting an intuitive impression about international phenomenon. Through surveys of current work, we find that there are some defects in the constructing of international network which is the basis of related work. Current networks are generally constructed on structured data through manual or semi-manual methods. It

limits the source of data, consumes too much manpower and time, and also cannot ensure timeliness and consistency. In order to solve these problems, we design a method that can construct international network from unstructured data of texts. The method achieves the goal to recognize countries and their relations automatically by analyzing texts. It expands the data source of international network greatly and also provides an efficient way of acquiring knowledge from information which can be used to solve the problem of information explosion. It is well known that Internet has became the richest source of information with the development of network especially Web 2.0. Thousands pieces of news are published on Internet every day. Through the news, people can learn what happened in the world easier and more rapid than even before. Meanwhile, too much information also immerses people in the ocean of information. People can hardly read all news and gain a comprehensive view about the world. It is a classic difficulty of the information that can be expressed as cannot see the wood for the trees. The international network transfers various events among countries into different relations and exhibits them in the form of visual graph. Graph has the feature of intuitive so that people can grasp the overview of international phenomena at a first glance. With the help of visual graph, people can also find hot spots and search what they are interested in easily. Overall, the labyrinth of international phenomena becomes clearer and easier to understand through international network.

The major innovations of this thesis can be summarized as follows.

(1) The method of constructing international network. As there has been no way which can construct international network from text automatically in current work, we originally propose our method of construct international network from large scale texts through text mining technology. It expands data source from structured data to unstructured texts and provide an efficient way to solve the problem of information explosion. We also build a comprehensive system including data acquisition, network construction and visual interface

to realize our method. Through experiment, we verify the reliability of our system.

(2) Definition and extraction of international relations. Based on related research about social network and interpersonal relationship, we point that intensity, quality and status are three important features of international relations. They affect the structure and evolution of networks in different way. We design several methods to extract these features from texts and construct the international relation network. Results of experiments show satisfied performance. Finally, we attempt to propose a new frame of defining international relations which contains five new types of relations based on the three features.

(3) Recognizing sentiment of relations between entities. Sentiment between countries reflects quality of relations which is an important feature of international relations. Although some work has been done on sentiment analysis, there is little work about recognizing sentiment between entities from texts especially in Chinese. A method composed of three steps is proposed. Through entities recognition and extraction, sentiment related region detection and sentiment determination; we can obtain sentiment between entities on sentence level. We compare different algorithm based on different principle (rule and machine learning) and different related region as well. The algorithm using CRF (conditional random fields) model based on syntactic dependency tree acquires best result.

(4) Design and realization of visual interface. We analyze defects in current visualization approaches of international relation network and propose a new method combined with GIS (Geographic Information System). The visual interface uses Google Maps as substrate, integrates Google Maps API and Mysql database and achieves international relation network visualization on electronic maps. The interface also provides some research functions that allow users to search interesting details and see different aspects of the international network more clearly.

(5) International network analysis. We analyze the constructed international network

through visual interface and obtain some interesting discoveries about hot spots and characters of network structure.

KEY WORDS: International Network; Extracting Relation; Features of Relation; Sentiment Analysis; Visual Interface

Chapter 1 Introduction

1.1 Background and Motivation

The current age is characterized by networks. With the development of Internet, people all over the world can conduct communications with others easily and form various kinds of networks. Various news are uploaded to Internet all the time, even a trifle happened in the corner can be spread all over the world in a short time. Actually, the whole world becomes a huge network and with closer relations than ever before [1]. Therefore in this paper, we propose a special network based on news from Internet that is international networks composed of countries and their relations. It can be used to extract important information from ocean of news and study the international phenomenon [2].

A network is a set of units (nerves, individuals, institutions, states), and a rule that defines whether, how, and to what extent any two units are tied to each other. According to the definition, an international network is also composed of nodes (countries) and connections (relations) among these nodes [3, 4]. If communications among countries can be transformed into connections of networks, the international phenomena can be represented in the form of international network. Through introducing method of network analysis, people can study the international phenomena from the view of network. For example, from the structure of the international network, people can identify status of each country. As actions of nodes in network are affected by their relations significantly, international networks can be used to study international relation evolution as well [5,6].

The existing of international network is with the increasing of international

communication and explosion of information. In ancient time, international communication is few and always limit in neighbors. International relations are simple and easy to handle at that time. With the development of communications technology, people have the ability of knowing what happened all over the world. World has really became a huge network. In recent years, as we all know, information on Internet is exploded in an incredible speed. People can hardly read all news in the world. It is an interesting paradox. On the one hand, various information seems can help us studying international phenomenon easier than ever before. On the other hand, people are immersed in the information ocean. They always cannot see the wood for the trees and cannot find the information they need [7]. International network is a proper way to solve the problem. As it transfers events among countries into different relations and exhibits them in the form of visual graph, people can get the overview of international phenomena easily [8, 9]. The form of network also allowed people to study international phenomena systematically through method of network analysis.

1.2 Research State of International Network

Recently, international networks are used in such research fields: cooperation and conflict among countries, terrorist network, international governance and economic. The approach of networks provides an efficient tool to describe, analyze and evaluate relations in these fields [10-13]. Through network, complicated relations become clear and visual. Researchers can systematically study not only relations but also their structures and effects. Generally, research of international networks follows such steps: decide the aim of network, collect data, define nodes and relations in networks, extract relations and construct networks, analyze networks and derive conclusions, show networks and conclusions.

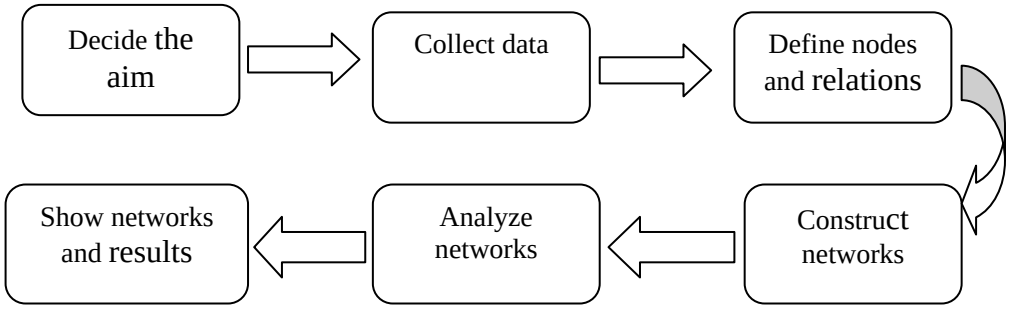


Fig 1-1: The general algorithm of international network research

In the processing above, collecting proper data and defining relations are basis of international networks. The term of “garbage in, garbage out” shows the importance of data for analysis. If the data is not comprehensive or out of date, the value of result will be reduced greatly [14, 15]. Relation is the core of networks. A network cannot be constructed without relations and relations decide types of network. However, current research has some defects in these two crucial steps. The principal methods of constructing international network are based on structured or semi-structured data including notice of international conference, trade statistics database, member list of international organization. It is not a hard work of constructing international network from such data. Take “The Correlates of War Diplomatic Exchange Data Set” as an example [16]. This database is about the relation of countries in the war. It describes relations between two countries in the war from 1817 to 2005. In the database, such parameters are provided:

Code 1- country1 of the war

Code 2- country2 of the war

Year- time of the war

DR_at_1-diplomatic relation of counry1 to country2

DR_at_2-diplomatic relation of counry2 to country1

DE – whether there is a deputy or not

Version- version of database

There are totally five symbols (0-3,9) to represent diplomatic relation from none to the ambassador level.

The international network can be obtained easily from the database if we treat codes as nodes and DR as connections. The five symbols of relation provide natural classification of relation. Constructing international network from such database has the advantage of simple and easy as database is well organized with clear classification of relation. International network based on database are generally easier to be accepted due to the authority of database as well.

However, such method also has obvious disadvantages. First, researchers should search proper database which is proper for their studying. It is an impossible mission sometimes as special databases for international relation are not much. A research about diplomatic relation network in East Asia had to use indirect data such as number of air route; freight turnover for there is no direct database about diplomatic relation. Second, as the databases are built by human beforehand, it always cost too much time and manpower. Research on these data cannot catch up with the rapid changes of international situation comparing with news on Internet. At last, there are missing data or inconformity in database, it affects further research. Comparing with the limited database, news on Internet can be regarded as inexhaustible resources. With features of comprehensive and timely, news is a perfect resource to construct international relation network.

Besides data, current international networks lack proper definition of relations. Relations are not classified or are simply separated into positive and negative. This defect limits in-depth research, as recent research has revealed that relations are more complicated than two types and different relations will lead different behavior in a network. Besides this, a vivid interface is a necessary part of international networks. Through the interface with ability of human-computer interaction, international networks become more clear and easy to

grasp. Reviewing related work, few researchers can provide such an interface. In the next section, we will describe related work and discuss their disadvantages in detail.

1.3 Key Points of Research

In this thesis, we tried to propose some methods to solve the problems mentioned in section1.2, our main achievement can be summarized as followed:

(1) We present an integral system that can construct international network from news and expand data from structured database to free text. We verify it through experiments and comparison.

(2) We find that intensity, quality and status are three important features of international relations. They affect the structure and evolution of networks in different way. Different methods are designed to extract these features from texts level to sentence level.

(3) Recognizing sentiment of relations between entities. Sentiment between countries reflects quality of relations which is an important feature of international relations. Although some work has been done on sentiment analysis, there is little work about recognizing sentiment between entities from texts especially in Chinese. We discuss how to obtain sentiment between entities in sentence level. Comparing different algorithm based on different principle (rule and machine learning) and different related region, we find that CRF (conditional random fields) model based on syntactic dependency tree acquires best result.

(4) A visual interface based on google map is developed. It integrates Google Maps API and Mysql database and achieves international relation network visualization on electronic maps. The interface also provides some research functions that allow users to search interesting details and see different aspects of the international network more clearly. Through analysis on constructed international network, we obtain some interesting discoveries about hot spots and characters of network structure.

1.4 Dissertation Outline

The rest of the thesis is organized as follows: we review the related work in Chapter 2, and in Chapter 3 we introduce a prototype system of constructing international network and evaluate its performance. Based on the prototype system, an improved system is proposed in Chapter 4. Chapter 5 describes a visual interface that can show the international network on world map. Section 6 concludes this thesis and gives some future work.

Chapter 2 Related work

2.1 Overview of International network

International network is a formalization method to study international relation. It regards entries (countries, organizations) as nodes, relationship (diplomatic relation, trade relation) as connections. The complicated international relation could be represented by a network. Through introducing methods of network analysis, researchers can study the evolution of international relation. In this section, we will introduce related research about international network and find some aspects could be improved.

International network is developing with communication and information technology. With convenient communications, people can know what happened all over the world. The world becomes so inseparable that a butterfly flapping its wings in the Amazon rainforest can generate a violent storm on the other side of the earth. Actually, countries in the world have formed a huge network and affect each other. In such a case, people hope to promote international governance, collective cooperation under the structure of network. Although researchers in international politics have realized the world became a network and used the conception of network to discuss some affairs, they did not there is a method of network analysis could be used to study the network of international relation for a long time. It is no strange as network analysis was first used in mathematics and computer science. Network analysis was not introduced into social science until the emerging of social network analysis.

International network analysis has close relation with network analysis. Network analysis is originated from 1930s, which is used to analysis the emotional relationship among people. In 1954, the name “social network’” was first proposed by Barnes White

studied the mathematical model of social network further form 1960s to 1970s. They proposed some important conceptions, such as strength vacancy chains and block models. Their research promoted the formalization of social network analysis. After that time, social network analysis had became an exact science [17-19]. Nearly the same time, some researchers began to introduce network analysis to international politics. In recent years, a series of achievement can be found in international network such as effect of network, the interaction of different network.

The overview of international network research could be referred in Networks of Nations: The Evolution, Structure, and Effects of International Networks, 1816-2001 [4]. It is the first book using network analysis approach in the study of international politics relations. In this book, Zeev Maoz combines social network analysis and world politics under the frame of international network to discuss evolution and change in the world system. He proposes a view that international relations are about networks, most interactions among states take places within different networks. Maoz believes that international network determines how information and influence flows in the global village. The networks also can help us understand international phenomena from Wall Street collapse to Persian Gulf War.

The book offers a systematic description of the evolution of international relations as a system of networks. In order to verify its theory of networked international politics, Maoz builds four types of international network for analysis which are alliance, trade, IGOs and diplomacy. He discusses structure, formation and evolution of these networks through four network indicators which are the density, polarization, interdependence, and transitivity. Anyway, the work of Maoz proves that network approach is a suitable way to the study of international politics. There are also some disadvantages in the work. The data used to construct international networks are come from various datasets such as the Issue Correlates of War dataset, Alliance Treaty Obligations and Provisions (ATOP) dataset. Data from

datasets has the feature of structured which are clear and easy for future processing. However, constructing datasets cost much time and human. Generally, it cannot be provided in real time. Therefore international network based on datasets cannot catch up with current international situation and help decision making on time. Data integrity is also a problem of dataset. Some missing data will affect analysis result. The missing data of trade is even up to 49%, which results query about Maoz's work. Besides that, the exhibition form of international work also has some disadvantages.

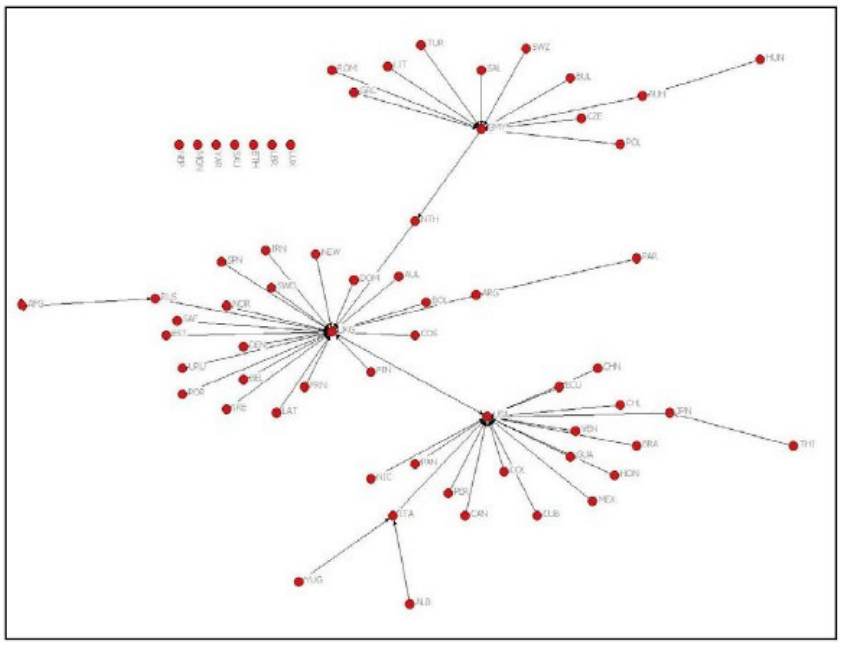


Fig 2-1: An international network about trade [4]

Fig 2-1 is an international network about trade in Maoz's work. In this graph, nodes are countries and lines are their trade relation. The isolate nodes on the left are due to lack of data. We can obtain an intuitive impression of world trade from this network. Through the network, stream of trade could be traced. Three main trade centers existing in the network which are America, England and Germany. Among them, Germany is connected to England through Netherland while America and England are connected directly. The relation between America and England is extremely close. This network can explain why the collapse of Wall

Street affects England first and Germany later from the side.

Although this network shows international trade vividly, the way of exhibition is not very satisfied. Countries are abstracted into little spots without their true locations. We cannot recognize which countries are there are without seeing the names carefully. This form of expression is suitable for common network but not so good for international network. An important characteristic of international network is that it composed of countries, and countries have their true geographic positions. The positions may affect network as well. For example, important countries in trade network are always located in crossroad. So, we should not ignore positions of countries when visualize international network.

2.2 Constructing International Network from News

In recently years, Internet has become the richest information source. Thousands of news about international phenomena are published on Internet every day [20]. If researchers can use such news to construct international network, the range and time effectiveness could be improved efficiently.

Recently attempt about building international network from news can be found in the research of Hämmerli [21]. He constructed a network about conflict in Chechnya based on 2818 events of news collected from 2002 to 2005. The network is composed of 44 actors and their relations. Actors are considered as nodes while events as relations. Different values from -13 to 7 are assigned to the events in order to reflect the relation of conflict and cooperation. Through this network, they identify main actors of conflict and cooperation that provide evidence to future solution. The result of network analysis is similar to that of human experts. This work proves constructing international network from new is reasonable and valid. It is a pity that the network is constructed manually. Human identify the actors and assign value to relations through reading events. That is not surprising as Hämmerli is a

political scientist. His main focus is not in constructing network automatically. Without doubt, this work should be carried by researchers in information science.

Although there is few works about constructing international network automatically from news, similar work about extracting relations among people and constructing social network can give us some useful experience. Some work can be traced back early to 1990s. Kautz tried to extract relation from text on web and construct social network. They defined relation between two people if their names appear on the same web. Jaccard coefficient is used to evaluate the intensity of relation. As an exploratory work, the system seemed a little simple [22].

Matsuo el.al proposed a system called POLYPHONET to search relations among researchers through search engine. Through query names of researchers, POLYPHONET can obtain occurrence of individual researcher, co-occurrence of two researchers and web about the two researchers. Occurrence and co-occurrence are used to calculate value which determinate whether researchers has relation or not. If two researchers have relation, web about them are analyzed to find types of relation. Four types of relations are defined beforehand (co-author, co-lab, co-project and co-conference).They used C4.5 classifier based on six keywords sets to decide types of relation. The system constructed a social network including 503 researchers and achieved satisfactory performance [23].

Yang had tired find social network on Chinese news. They searched main verbs and relative entities in sentences. Entities are nodes in the network while verbs represent the relation. A social network with positive and negative relation based on 50 pieces of news was proposed in their work [24].

From their work, we can get the basic steps of constructing network from text. As networks are composed of nodes and their connections, the task of constructing network from news means identifying entities and their relations automatically. Identifying entities is

a relative easy task with the development of named entity recognition. By comparison, extracting relations is much harder. On the discussion above, we introduce methods about extracting relations on texts. Recently, with the development of social network websites, researchers of social networks pay attention to extract different relations from these websites as well. They always extract relations from macro features, such as activity and interaction of entities. Marlow extracted three types of relations among users on Facebook [25]. The types of relations are depended on different interactions of users, such sending, reading, replying message. Kazienko studied direct and indirect connections from photo sharing system and discussed the effect of different relations in recommender system [26]. Jerome et al. detected positive and negative edges through tags of friends or foes based on Slashdot Zoo corpus [27]. As macro features are about activity and interaction, the relations based on macro features are more useful for future work. Take Jerome's work as an example, They not only extract relation ,construct social network but also analyze three levels of characteristics on the network and verify multiplicative transitivity that can be summarized as the enemy of my enemy is my friend on these three levels.

Introducing macro features to extract relations of international networks is a proper method. However, the researches mentioned above extract relations from data about users' activities that collected by website. For news, there is no such direct information. We should design method that can extract macro features of relations from news.

2.3 Visual Interface

In our investigation, the e-diplomacy Hub released by Agence France Presse (AFP) in 2012 is the best system of showing international network so far. The main interface of the system is shown in Fig 2-2. In this interface, users can select one country to search its e-diplomatic network, the other nations with which it communicates, and the volume of

communication. It also allows adding another to see their relation. Here, we choose China and Japan. Green line is the relation network of China while red line is Japan's. The shade of line indicates volume of communication. Users can click the mark on countries to see related information. In the corner of the interface, hotspots are shown as well [28].

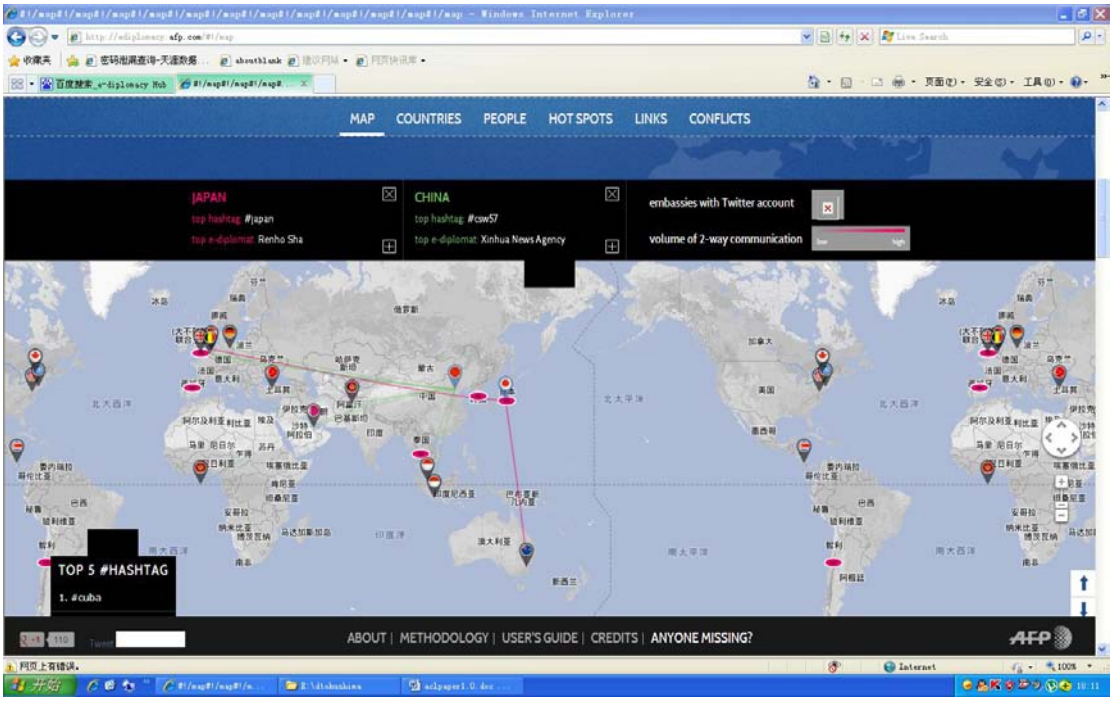


Fig 2-2: The e-diplomacy Hub system

The e-diplomacy Hub is based on data that collect from thousands of tweet accounts about diplomacy every day. It achieves provides international network in real time. The ranking of countries and their interactions are obtained through calculation of the real time data. Through the map interface, users can see hotspot in the country today and search countries their interested in. Although the system can analyze, measure and visualize international relations in real time, it still lacks ability of showing international network. The system only provides one type of relations without classification. In the interface, users can only search relations of one or two countries and cannot obtain the whole international network.

Anyway, disadvantages of current work can be summarized as :

1 Proper data source

2 Proper way of defining and extracting relations.

3 A suitable interface

We try to improve these disadvantages in our system. The system has new definition of relations and can use news as data resource and extract relations automatically. Therefore, an International networks can be provided in real time. With the visual interface of rich functions, users can search interested detail and carry further research.

Chapter 3 Prototype System

3.1 Introduction

3.1.1 General Algorithm

Based on related work in Chapter 2, we propose a prototype system which can construct international network from news. The general procedure is as followed: first collect news about international relation, carry some necessary preprocessing on the news, such as word segmentation. Then we extract countries from news as nodes. After detecting nodes, we would decide which countries have relations and extract these relations, give relations proper classification. International network would be constructed on these nodes and relations at last. The algorithm is shown in Fig 3-1.

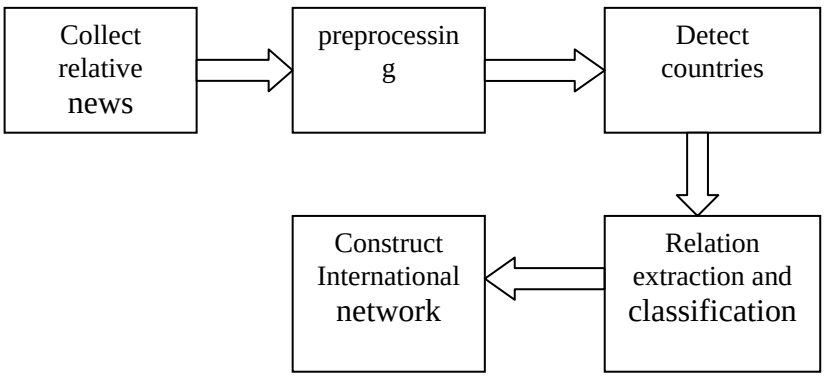


Fig 3-1: The algorithm of constructing international network

The most important and difficult work in the algorithm is relation extraction and classification. In previous work of international networks, researchers focus only on whether the relations existing or not and ignore precise definition of different relations. Relations are not classified or separated into conflict and cooperation simply. In fact, international relation

is more complicated than conflict and cooperation. Take relations among America, Iraq and Cuba as an example. Although both of the latter are enemies of the former, there are some differences. America directly involved in the war against Iraq. By contrast, America did not want to have any relation with Cuba, neither war nor friendly communication, just blockade, which is slightly like the cold war. As these two kinds of conflicts yield different effects and actions, they must be defined as different types of relations. Here, we refer some work about social network analysis and interpersonal relationship.

3.1.2 Relation in Social Network Analysis

Researchers of social network analysis noticed that different relations play different roles in network structure and information transmission. Intensity of relation was analyzed first. Granovetter analyzed the social network about job-hunting in 1960s, he found that most people find job with the help of personal relationship which means social network playing important role in job-hunting. However, the most helpful relation is not close friends but ordinary friend. Granovetter called it weak ties which is more efficient in information transmission. In his research, Granovetter proposed four features to evaluate intensity of relation as well [29, 30]. Besides intensity, social network also pay attention to quality of relation which means positive or negative relation [31]. In traditional social network, all relation was considered as positive relation that represents cooperation. However negative relation in practice is not rare. How to represent these negative relation and analysis their effect in network became hot issues recently. Jerome's work about positive and negative edges on Slashdot Zoo corpus is a good example [27].

3.1.3 Relation in Interpersonal Relationship

Research about interpersonal relationship can provide some useful experience as well. Interpersonal relationship is the relation between human while international relation is the

relation between countries. Form a macro-level, all the world is a global village, countries are villagers. International relation is similar to interpersonal relationship in some ways.

The most widely accepted research about interpersonal relationship up to now is proposed by American psychologist A.Lewicki. Lewicki analyzed interpersonal relationship between couples through an experiment which looks like a game of 90 minutes [32]. Based on research over 1000 couples, Lewicki divides interpersonal relationship into eight categories. They are subordination, cooperation, competition, subordination-competition, subordination-cooperation, competition-cooperation, subordination-cooperation-competition and random. We can see that there are three basic relations, i.e., subordination, cooperation and competition. The interpretation of three relation is as followed:

Cooperation: couple has a consistent goal; they work as one for the goal. Furthermore, they have equal status while working.

Subordination: Although couple in this relation also works for a consistent goal; their status is different. One is in the dominant position and the other is in the subordinate position.

Competition: This relation has features of quarreling and competition. Couple has their own goals and blame each other.

We can see that in Lewicki's work, they classified interpersonal relationship with features of quality (cooperation or conflict) and status (equal or unequal). This classification is reasonable in relationship like couple. However it has a little defect that ignores intensity of relation. Research on couples has a premise that people have a relation of conjugal. This relation has features of being steady and intimate as it is established and protected by law. Couples cannot live as strangers or enemies unless they divorce. Comparing with interpersonal relationship, Lewicki's work ignored relations like strangers and enemies.

3.2 International Relation

From the discussion above, we find that in research about social network and interpersonal relationship, researchers define relations on features like intensity, quality and status. We combine them under the unified framework in our work and propose five categories of international relations: confrontation, neutral, cooperation, subordination and conflict.

Confrontation: Two countries are estranged from each other. Although they both have their own external contacts and friendly nations, they do not like to contact with each other and evade co-occurrence in an international event. The association between them is minimal and they can be engaged in a cold war.

Neutral: Two countries have normal exchange, neither estranged nor close, which resembles the relation between two persons in a same community who just greet each other when meeting on road.

Cooperation: Two countries have a close and friendly relation. In addition, they hold equal status while getting along with each other.

Subordination: Although two countries have a close and friendly relation, their statuses are unequal. The relation is more important to one country than to the other.

Conflict: Different from friendly relationship, two countries in conflict have a close relation but it is represented as quarrel and war. The difference between confrontation and conflict is that the former means a cold war while the latter can mean an ongoing war.

From above, we see three main factors that determine the types of relations. As shown in Table 3.1, they are intensity, quality and status of relations.

Table 3.1 Factors of relation

Relation	Intensity	Quality	Status
Confrontation	estranged	null	null
Neutral	normal	null	null

	close	neutral	null
Cooperation	close	friendly	equal
Subordination	close	friendly	unequal
Conflict	close	conflict	null

Then we try to seek features from texts that can be used to represent the three factors. In the traditional social network, co-occurrence is usually used to estimate the intensity of a relation [33, 34]; in the existing research of international networks, transitive verbs are used to decide the quality of a relation; while subordination coefficient can reflect the status. We introduce them in our work with some improvement. The final three features we extract from texts and use are Chi test, subordination coefficient and sentiment value. The former two are macroscopic ones based on country co-occurrence at document level and the last one is a microscopic one based on the analysis of sentences. To the best of our knowledge, there is no method that can use features from a document level to sentence level. The details of these features are discussed next.

3.2.1 Chi Test

Chi test is one way of measuring the intensity of a relation [23]. Actually, several indices can be used in measuring it, such as matching coefficient, mutual information, Dice coefficient, Jaccard coefficient, overlap coefficient, cosine and Chi test [35]. Different methods can lead to different results. For example, matching coefficient considers only times of co-occurrences while overlap coefficient measures co-occurrence through inclusion. We choose Chi test because it fits our requirement to distinguish relations. It is defined as follows:

$$V_{Chi} = \frac{(ad - bc)}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} \tag{3-1}$$

In the equation, *a* denotes the number of events containing both two countries (*c1* and *c2*), *b* and *c* denote the number of events containing only one country (*c1* or *c2*), *d* denotes the

number of events without these two countries (neither *c1* nor *c2*). The range of value V_{Chi} is from -1 to 1, where a positive value means positive association while a negative one means negative association. The greater the absolute value, the stronger the correlation [36].

Chi value is widely used in ecology to measure the relation between two species [37]. If two species have symbiotic relationship and often appear with each other, a higher value is assigned, which indicates a positive relation; while if they have the relation of competition and often appear individually rather than co-appear, a lower value is obtained, which indicates the exclusion of each other. It can be introduced into an international relation as well. If two countries have a close relation, they are willing to communicate with each other, and thus high frequency of co-occurrence can be found. Conversely, confrontation leads to few exchanges that can be revealed by a low Chi value.

3.2.2 Sentiment Analysis

Sentiment analysis is used to detect the quality of relations. It is a microscopic feature at a sentence level. Through a sentence containing countries, we should evaluate if the relations among them are positive or negative. The related research has a long history and existing methods are based on transitive verbs generally. The initial effort can be traced back to the earlier 1990s. Schrodt et al. propose their KEDS (Kansas Event Data System) [38]. Based on the fact that most events are defined by sets of transitive verbs (verbs that have a direct object), they develop the program that can identify the basic SVO (subject-verb-object) structure of an English sentence. While the SVO structure represents the event, the transitive verb determines the event code, which can be seen as the sentiment between a source and target. Due to the limitation of NLP (natural language processing) technology at that time, KEDS suffers from many problems such as passive voice and ambiguous words. Later some systems followed similar method. For example, in IDEA (Integrated Data for Events

Analysis) system, the form of "who does what to whom" extracted from a clause is used to represent the event. "What" is the sentiment between "who" and "whom"[39, 40]. Yang had tried this method in Chinese as well to find positive and negative relations in a sentence [24]. Although dealing with only transitive verbs in a clause is simple and convenient, such methods have some problems especially when dealing with Chinese that is more flexible than English. First they cannot deal with more than two entities or entities without transitive verbs. Take the sentence (“The allied forces of America and Britain attack Iraq.") as an example. Three countries appear in the sentence and form three pairs of relation. America and Britain are ally with a positive relation while Iraq is their enemy with a negative relation. The methods based on transitive verbs cannot identify the relation between America and Britain that are connected with no transitive verb. Second, although verbs play an important role in expressing sentiment, other words cannot be ignored. Based on the analysis of Chinese emotion keywords, about 40% of emotion keywords are verbs, while the rest are composed of nouns, adjectives and so on. Thus we cannot simply focus on transitive verbs only [41, 42]. Most current machine coding systems have similar problems with a varying degree. The lack of Chinese emotion dictionary affects sentiment analysis in Chinese as well.

For the first problem, recent advance of sentiment analysis in both English and Chinese give us some useful ideas [43-45]. The third Chinese opinion analysis evaluation (COAE2011) sets a task about recognizing different sentiment of different attributes in one sentence. Most teams search the related region of different attributes first. The following sentiment analysis is based on the region. Similar method can be found in English as well. Kim and Hovy find the related regions of opinion in the sentence based on holder and target before determining the sentiment of opinions [46, 47]. Note that the aim of sentiment analysis is to identify the opinion of a target while our task is to detect the quality of relations. The principle of detecting the related region can improve our work as well. We can

also find the related region of different country pairs first, and then perform further processing on the region. The method based on a syntactic dependency tree is introduced to detect the related region of country pairs. A sentence after parsing forms a syntactic dependency tree. It reveals the syntactic component of words and dependency relations between them [48, 49]. For that reason, one can find the region that has syntactic relevance with country pair. In our method, the related region is the dependency chain that connects two countries in a dependency tree. For different country pairs, the related regions are different.

For the second problem, we expand the range of emotion word with help of our previous work in the field of sentiment analysis [50, 51]. In our method, all emotion words in the dependency chain (not only verb) decide the sentiment attribute among entities. In order to analyze sentiment, an emotion corpus must be built beforehand. We build it based on former work [41]. Through the analysis of 500 Chinese blogs from July 2008 to January 2009, more than 10000 Chinese emotion keywords are annotated and separated into eight classes (expectancy, joy, love, surprise, anxiety, sorrow, anger and hate) with different values from 0 to 1 by hand. The greater the emotion intensity, the bigger the value. As only positive and negative sentiment is under consideration in our work, we reduce the eight classes into two. Expectancy, joy and love are classified as being positive; anxiety, sorrow, anger and hate are classified as being negative and surprise is abandoned. For example, "喜欢(like)" (expectancy=0.0, joy=0.3, love=0.9, surprise=0.0, anxiety=0.0, sorrow=0.0, angry=0.0, and hate=0.0) has positive sentiment value of 0.4((0.0+0.3+0.9)/3=0.4). We also add some words to the corpus manually. After all, there are 15913 Chinese words in the emotion corpus.

In brief, the proposed method of sentiment analysis can be summarized as follows. First, find country pairs in a sentence. Then, search related regions of these pairs through the syntactic dependency tree. From the sentiment of related regions, the qualities of relations

are obtained at last.

3.2.3 Subordination Coefficient

Subordination coefficient is used to distinguish the relation of cooperation and subordination. It is a macroscopic feature at a document level. Although cooperation and subordination are both friendly relations, they are different. In cooperation, two countries have same importance toward each other while in subordination, one country relies more on the other. If a bilateral relation plays an important role in diplomacy of one country but not so important to the other one, we assert that they have the relation of subordination. For example, country A and B contact each other five times in one year. Country A has a total foreign intercommunication of ten times in the year while country B has one hundred times. Obviously, their relation has the feature of subordination and the bilateral relation is more important to country A. Subordination coefficient is defined as follows:

$$V_{subordination} = \frac{n_{c1 \cap c2}}{n_{c1}} - \frac{n_{c1 \cap c2}}{n_{c2}} \tag{3-2}$$

where $n_{c1 \cap C2}$ denotes the number of events containing both two countries (*c1* and *c2*). n_{c1} denotes the number of events containing *c1* and n_{c2} denotes the number of events containing *c2*. Obviously, $V_{subordination}$ has the value from -1 to 1. The greater the absolute value, the stronger the subordination. Based on the above discussion, we summarize the features of five types of relations in Table 3.2.

Table 3.2 Features of relation

	Chi	Sentiment	Subordination
Confrontation	negative	*	*
Neutral	Close to zero	*	*
	positive	neutral	*
Cooperation	positive	positive	Less than threshold
Subordination	positive	positive	Larger than threshold
Conflict	positive	negative	*

* means that we do not consider this feature in the corresponding relation

3.3 Algorithm and Data Processing

3.3.1 Data Source

It is not an easy task to search suitable corpus for the analysis of international relations as it requires large-scale and continuous data about relations among countries. Data about international relations on Internet are generally scattered in the flow of international news. We tried international news posted at sites such as sohu (a popular Chinese website offering the search service). However, most of them are economic news related to companies or non-governmental organization (NGO), not countries. Nowadays, some researchers are willing to use the data collected by governmental or professional institutions in order to ensure authority and continuity [52, 53]. We choose current affair-related data from the magazine called China Comment. It is sponsored by Xinhua news agency (the most authoritative one in China). Every year, it issues important current affairs happening overseas in the current year [54]. The data are divided by the month with tens of items of events every month. Each event is described briefly with the length of hundreds of Chinese words. It has been in existence for more than 20 years and described the world from the view point of China.

3.3.2 Algorithm Description

We collect the current affairs of 2009 (210 pieces) and 2010 (173 pieces) as the input. As they are stored in one text document, some data preprocessing must be performed. The document should be divided into pieces; each piece corresponds to one event. Then, country names in each piece are extracted based on the list of countries all over the world. The list contains more than 200 countries; it also includes United Nations as the most important

international organization in the world. Some events have the problem of co-reference resolution as countries may have different names in the corpus. For example, Britain can also be expressed as the United Kingdom. We use a rule-based method to unify names representing the same countries, establish the mapping of a country name with its abbreviations and alternative names manually.

After preprocessing, we establish pairs of countries based on co-occurrence to determine whether they have a relation or not [55]. If two countries co-occur in more than one event, we assume that they have a relation and continue to extract their features of Chi value, subordination coefficient and sentiment. The extraction of three features is conducted in parallel.

To calculate Chi value and subordination coefficient, countries that appear in a same event are deemed to have a co-occurrence. After counting co-occurrences and occurrences of countries, we can calculate Chi value and subordination coefficient of countries according to Eq. (3-1) and Eq. (3-2). It should be noted that in calculating Chi value, *b* and *c* are different from *nc1* and *nc2* in Eq. (3-2). The latter are the numbers of events that contain *c1* or *c2* while *b* and *c* are the numbers of events that contains only *c1* or *c2*. In 2009, the total number of events is 210, and

$$d = 210 - a - b - c \tag{3-3}$$

For sentiment analysis, we follow the steps of finding sentences with country pairs, searching related region and detecting sentiment. In the first step, we search each document to obtain sentences that contain at least one country pair. In some documents, two countries do not appear in the same sentence even though they appear in the same document. For these documents, the processing of sentiment analysis is skipped.

The second step is to search the related region based on a syntactic dependency tree. After obtaining sentences and country pairs, we use Chinese parsing tool of LTP (Language

Technology Platform) to form a dependency tree and find the related region [56]. The related region is a dependency chain in the tree that connects two countries. If the dependency chain contains main verbs, the sentiment of main verb represents the relation between countries, or else the sentiment of the word that attaches them represents the relation. Take the example of "美国和英国两国联军出兵伊拉克。(The allied forces of America and Britain attack Iraq.)". The sentence after parsing is shown as below:

美国|ATT,4 和|LAD,4 英国|ATT,3 两国|ATT,4 联军|SBV,5 出兵|HED,-1 伊拉克|VOB,5

Here, English abbreviation is a constituent part of the word in this sentence; and the number is what the word relies on. "HED,-1" means that the word is the main verb of this sentence. It forms a syntactic tree in Fig. 3-2. From the tree, we can easily see the dependency chain among countries.

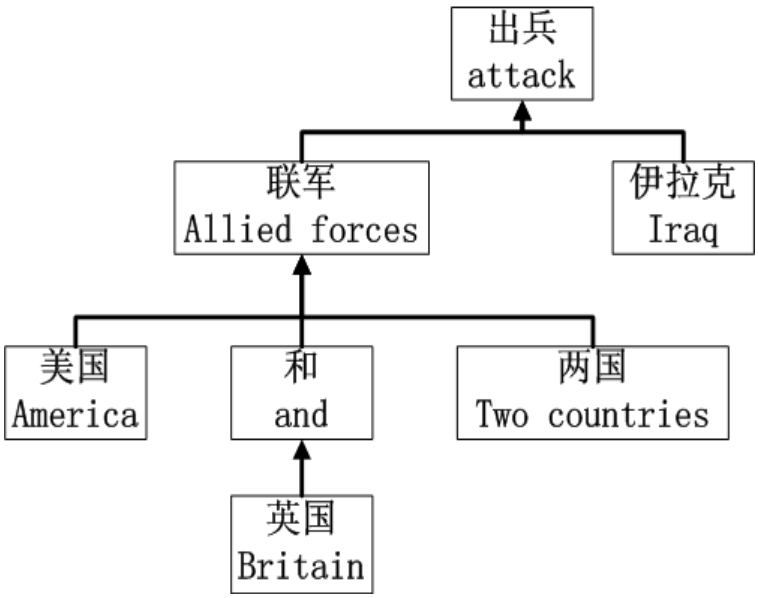


Fig 3-2: Syntactic dependency tree

Two pairs of countries (America and Iraq, British and Iraq) have a dependency chain with main verb "出兵(attack)" while America and British have a dependency chain with a connecting word "联军(allied force)". In the last step, from the emotion dictionary, we can

obtain sentiment values of words “出兵(attack)” and “联军(allied force)” which are used to represent the relations of these two pairs.

The pairs of countries with a Chi value, subordination coefficient and sentiment value are stored in the same database. When all three features are extracted, the final result of categories can be easily obtained according to Table 3.1. The algorithm is shown in Fig. 3-3:

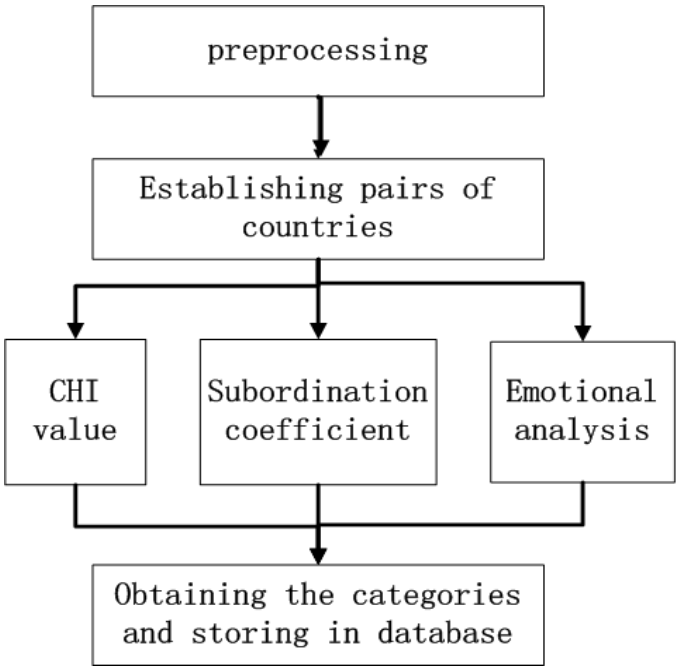


Fig 3-3: Algorithm to obtain the categories of relations among countries

3.4 Result and Analysis

We use the current affair of 2009 and 2010 from China Comment for the experiment. There are 210 events with 93 countries in 2009 and 173 events with 84 countries in 2010. Top ten countries that have the highest frequency in 2009-2010 are shown in Table 3.3.

Table 3.3 Frequency of countries

Country2009	Frequency	Country2010	Frequency
America	66	America	42
Japan	28	Japan	26
Russia	21	North Korea	23
UN	19	South Korea	19
South Korea	19	Russia	17
China	19	China	16
North Korea	15	Afghanistan	11

Afghanistan	13	UN	11
British	12	India	10
India	11	France	9

Frequency reflects the importance of countries. A country's name appearing more often means that it is more important in the world. It is not a surprise that America appears the most and it is indeed the most powerful country. Results of the two years are similar. Countries in the table are either important countries as the permanent members of UN Security Council or hot spot countries such as Afghanistan. On the other hand, the result of frequency distribution in accord with common view verifies the reliability of data source we choose. Comparing with Afghanistan, Iraq only appears once. It implies that in 2009, Iraq was no longer the focus of world. The rise of Korea's appearance count in 2010 indicates that there was something in the Korean peninsula.

As mentioned before, we assume that two countries have certain relation and form a pair if they show co-occurrence in more than one event (a piece of document about the event). Note that ignoring those pairs having only one co-occurrence in the dataset of a year does not affect the analysis result while allows one to significantly reduce the analysis complexity. In the 2009 dataset, we have found 170 such pairs for further processing and 130 ones in the 2010 dataset. To consider the relation of subordination, relation in a pair should be directed, i.e., China-America is different from America-China.

From country pairs, we can obtain the relation range of a country with others. For example, if there are ten pairs containing country X, then X has a relation with other five countries (notice that China-America and America-China are two pairs). Hence, X's relation range is 5. The result of relation ranges is summarized in Table 3.4 for 2009 and 2010. Russia ranks the top followed by America in both years.

Table 3.4 Relation ranges of countries where N=Relation Range

2009	N	2010	N
Russia	21	Russia	19
America	14	America	13

China	12	China	10
Japan	11	Australia	10
UN	10	Japan	9
India	7	UN	8
South Korea	7	India	8
Australia	6	France	8
British	5	British	7
France	3	Canada	7

Some countries with high frequency of occurrence did not appear in the table, such as North Korea and Afghanistan. Although they appear frequently in events, their external contacts are quite limited. Both countries had the relation with only three countries in 2009 and 2010.

3.4.1 Chi Value

Chi value is the measure of co-occurrence intensity. High positive Chi value means a close relation while negative value means confrontation. The following analysis is based on the 2009 dataset. As shown in Fig. 3-4, in all 170 pairs, only 12 pairs have the values below zero, while all the others have positive values. Even South Korea and North Korea have Chi value of more than 0.5 with the co-occurrence of 9 times.

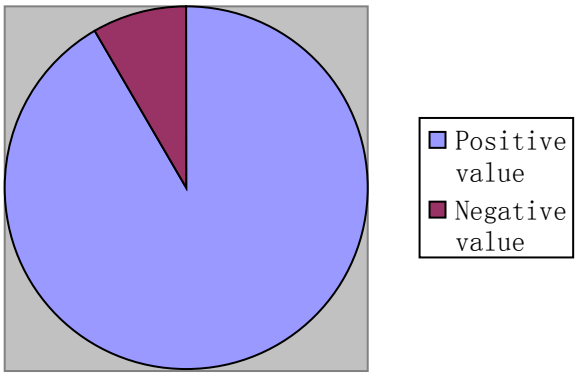


Fig 3-4: Result of Chi value

High Chi values often appear between neighboring countries that have natural connection such as Caucasus countries and Pacific countries. Most of them are not big powers while big ones have relatively low and average Chi values. It hints that big powers intend to

communicate with more friends rather than one close friend. Most low Chi values concern America. That is due to the high frequency of occurrences and low frequency of co-occurrences. In other words, America often appears in world events but does not like discussing with friends, i.e., plays the role of unilateralism. The lowest value is between America and UN, only -0.1420 which shows the strong confrontational nature. Five pairs that have the highest and lowest Chi values are shown in Table 3.5.

Table 3.5 Result of Chi value 2009

Pair	Value	Pair	Value
Azerbaijan-Armenia	0.8639	America-UN	-0.1420
Moldova-Armenia	0.8145	America-Japan	-0.0845
Moldova-Belorussia	0.8145	America-India	-0.0671
Australia-New Zealand	0.7708	America-Russia	-0.0547
Kazakhstan-Uzbekistan	0.7036	Russia-Japan	-0.0374

It is slightly beyond expectation that America and Japan have negative Chi value. That is partly because of the data source provided by a Chinese news agency. Reviewing Eq. (3-1), Chi is affected by the frequency of co-occurrence and occurrence. The more appearing independently, the less co-occurrence and the lower Chi value. As data favor more Asian events, more events contain Japan and Asian countries rather than America are chosen. It causes high occurrence of Japan, but co-occurrence of Japan and America is less than the expected. The same situation can be found between America and South Korea with a low Chi value of 0.0725.

The results based on the 2010 dataset is more or less similar to those of the 2009 one as shown in Table 3.6. High Chi values still appear in neighboring countries. Three pairs have slightly negative Chi values and all of them concern America.

Table 3.6 Result of Chi value 2010

Pair	Value	Pair	Value
Australia-New Zealand	0.7003	America-N Korean	-0.0887
Spain-Portugal	0.5680	America-China	-0.0581
Lithuania - Ukraine	0.5259	America-S Korean	-0.0031
S Korean- N Korean	0.7100		
Germany - France	0.7036		

3.4.2 Sentiment Analysis

Firstly, we discuss the reliability of sentiment analysis. As human annotation costs too much, we only analyze the result of 2009 dataset. Sentiment analysis is performed at the level of a sentence. If two countries as a pair appear in the same sentence, we call it an instance of the pair. Sentiment value of a pair instance can be obtained through the algorithm discussed above. After summing up all the emotion values of a pair's instances, we have the sentiment value of the pair. Note that prior-defined pair in Section 3.3.2 is based on co-occurrence at the level of a document and an instance of a pair is based on co-occurrence at the level of a sentence. In the 2009 dataset, more than 500 instances are found; most of them are simple coordinate instances. For example, in sentence "NATO includes America, Canada, British, French and so on." The four countries compose 12 simple coordinate instances, there is no emotion word among them and thus show no sentiment. Sentiment analysis based on a dependency tree is not necessary for them. Hence, we remove these instances from the result of sentiment analysis. After such removal, 236 instances are used for analysis. By human annotation, 62 instances exhibit negative relations while the others positive ones. We compare the result of sentiment analysis automatically with human annotation. There are 188 correct ones in all 236 instances with the precision of 79.66%.

In instances of errors, about half are related to parser errors. As we use the whole sentence that is separated by a period but not clause as the input of the parser, sometimes the long sentences cannot be resolved correctly. No correct syntax tree means no correct dependency chain and no correct result. Future improvement on the parser side is expected to improve our result.

Through the further analysis of errors, we have found that performance varies from event to event. The results on some events are good while bad on others. However, errors of negative instances are significantly more than positive ones. Although only one quarter of

the instances are negative ones, about a half of errors belong to them. The comparison between them is shown in Table 3.7. Three metrics of evaluation are $V_{precision}$, V_{recall} and $F_{measure}$. Note that $F_{measure}$ depends on the other two metrics as follows:

$$F_{measure} = \frac{2 * V_{precision} * V_{recall}}{V_{precision} + V_{recall}} \tag{3-4}$$

Table 3.7 Comparison of negative and positive instances

	Precision	Recall	F-measure
Negative	64.52%	60.61%	62.50%
Positive	85.06%	87.06%	86.04%

The situation of negative instances is more complex than that of positive ones for the reason of privative, which is the main cause of bad performance in negative instances. For example, in the sentence "朝鲜绝不同意联合国该项决议. (North Korea will never agree the UN resolution)". The relation between North Korea and UN is not "同意 (agree)" but "绝不同意 (never agree)". Our future work will study the effect of privative and hope to improve the performance of recognizing a negative relation.

At last, we have final sentiment values of all pairs. In these 170 pairs, 78 pairs hold positive sentiment, 12 hold negative sentiment and the other 80 pairs do not show sentiment orientation. Top ten pairs exhibiting positive and negative sentiment are listed in Table 3.8.

Table 3.8 Result of sentiment analysis

Pairs of positive	Value	Pairs of negative	Value
Australia-New Zealand	1.156	America-North Korea	-0.760
China-UN	1.004	America-South Korea	-0.421
Japan-India	0.544	America-Japan	-0.175
British-Afghanistan	0.500	Ukraine-Russia	-0.147
Switzerland-UN	0.467	America-Turkey	-0.113

The result of positive pairs corresponds to common sense more or less while negative pairs show comparatively large deviation. Two pairs (America-South Korea and

America-Japan) are out of expectation. Take America-Japan as an example. There are six events containing both America and Japan, four of which do not imply any sentiment and two have positive sentiment. Even for these two events, one of them has no sentence that contains both America and Japan, which means that there is only one instance of the pair of America and Japan. Error of the only single instance leads to the erroneous final value. The situation of America and South Korea is similar. If there are more instances of country pairs, the performance can be improved. For the pair of South Korea and North Korea, 7 instances containing sentiment are found, 4 are positive and 3 are negative. Although there are some errors in sentiment analysis, the final result (0.233) still shows positive sentiment. Adopting large corpus can improve the final result as well.

3.4.3 Subordination Coefficient

The subordination coefficient ranges from -1 to 1. Positive means master, negative means attendant. Two countries in one pair have the same absolute value of subordination but with different signs. Top five pairs of subordination coefficients are given in Table 3.9.

Table 3.9 Result of subordination coefficient

Pairs 2009	Value	Pairs 2010	Value
America-Mexico	0.970	America-Iraq	0.905
America-Canada	0.955	China-Brazil	0.875
Japan-Somalia	0.929	Russia-Kazakhstan	0.823
Russia-Moldova	0.905	Ukraine-Lithuania	0.714
Russia-Armenia	0.857	America-Qatar	0.692

Most of these pairs are great power and their weaker neighbors. Japan- Somalia and China-Brazil are unexpected results. Few events may be their main cause. There are only two events about Somalia in the corpus and both of them are related with Japan, one is Japan sending a convoy fleet to Somalia and the other is Japan sending ground self-defense force to Somalia. Situation of China-Brazil is the same. Only two events contain Brazil and both of them contain China as well.

3.4.4 Final Result

After obtaining all the results of the three features, we separate pairs into five categories given in Table 3.10, which also includes the number of the relations in each category. In current research, the thresholds are given by human.

Table 3.10 Final result

Categories	standard	2009	2010
Confrontation	VChi <0	12	6
Neutral	0< VChi <0.1	90	78
	VChi≥0.1 and Vsentiment=0		
Conflict	VChi≥0.1 and Vsentiment<0	8	8
Subordination	VChi≥0.1 , Vsentiment≥ 0 and |Vsubordination |≥ 0.4	26	18
Cooperation	VChi≥0.1 , Vsentiment≥ 0 and |Vsubordination |<0.4	34	20

In order to evaluate the result of relation classification, we use human evaluation method. Two experts assign a score to each pair in 2009 based on 210 pieces of news individually, 5 means perfect and 1 means worst. The scores of two experts have high consistency, only 8 pairs with scores difference more two. As shown in Table 3.11, the average scores of both two experts are more than 3.5 and in all pairs only a few pairs with scores less than 2. It denotes our method of relation classification is reliable.

Table 3.11 Evaluation of result

	Expert1	Expert2
Average score	3.53	3.59
Confrontation	3.58	3.75
Neutral	3.23	3.45
Conflict	3.88	3.75
Subordination	3.65	3.62
Cooperation	4.16	3.88
Less than 2	28	22

In this chapter, we propose an original approach that can construct international network

from news. It uses three features (intensity, quality and status) from macroscopic features at the document level and microscopic features at the sentence level to detect these international relations among countries. Based on the extensive analysis of results, we found that most results are reasonable. There are still some places could be improved such as sentiment analysis and scale of corpus. We propose improved methods to obtain better performance of sentiment analysis especially in a negative sentiment case; and acquire news from Internet as data source of our system.

Chapter 4 Improved System

We introduce a prototype system of constructing international network form news in Chapter 3. Although we verify the general method is reasonable, there are still some defects in the system. We propose an improved system in this chapter. It improves previous system in two main respects. First, a new method of sentiment analysis is proposed to improve the performance, second news from Internet are collected as data source.

4.1 Sentiment Analysis between Entities

From discuss in Chapter 3, we see that sentiment between countries represents quality of relation. As an important feature of international relation, it affects the performance of system directly. However, the result of sentiment analysis did not satisfy especially in negative sentiment. We propose a method combined with machine learning method to improve sentiment analysis. The goal of the approach is to recognize sentiment of relations between entities (Entities are countries in current experiment) from related sentences in Chinese. As not all countries appearing in the sentence show sentiment orientation between each other, in our approach we divided sentiment polarity into three types which are positive, negative and neutral (We divided sentiment polarity into two types in Chapter 3, therefore work in this section is more complicated). In order to accomplish the task, we design the method of three steps which are sentence selection, related region detection and sentiment recognition.

4.1.1 Sentence Selection

In this step, sentence that contains country pairs are selected for future processing. First of

all, we split text into sentences by full stop. Compared with the method of splitting sentences into clauses by comma, full sentences have more country pairs which are more complicated for analysis. Then sentence that contains more than one country is selected with country pairs. For example, we get three pairs of countries in the selected sentence “美国和英国两国联军出兵伊拉克.(The allied forces of America and Britain attack Iraq.)” Pairs are <美国英国>（< America Britain>）, <美国伊拉克>（< America Iraq >） and <英国伊拉克>（<Britain Iraq >）. The remaining task is to recognize sentiment in these pairs.

Although in this step we recognize countries from a list of more than two hundred countries and religions, it still has the problem of co-reference resolution as countries may have different names in the sentences. For example, Britain may also be expressed as the United Kingdom. We use rule-based method to unify country names, establish the mapping between country names and their abbreviations or alternative names.

4.1.2 Related Region Detection

As several pairs of countries may appear in one sentence, we face a problem of determining the related region in the sentence that express sentiment of each pair. The sentiment related region can be one word or string of words. Two kinds of main approaches are used to determine the related region. One is called bag-of-words and the other is based on syntactic dependency structures. The former finds related region based on position of word in the sentence. It does not require a sentence parser and has the feature of being simple. In Kim’s work, they tried four variations of region size based on bag-of-words [57]. We used this simple method in our research as well. String of words between two countries is selected as related region of the country pair.

As bag-of-words method ignoring modifier relations between words ,some words between two countries may not have syntactic relation with the two countries. We propose

a method based on syntactic dependency structure and hope it could improve the performance. A sentence after parsing will form is a dependency tree. It reveals the syntactic component of words and dependency relations between them. We use the parser tool of LTP developed by HIT-SCIR for processing [56]. Take the sentence we mentioned before as an example: “美国和英国两国联军出兵伊拉克.(The allied forces of America and Britain attack Iraq.)”. The sentence after parsing can be referred in Chapter 3

Totally, we have chosen three windows of related region in this section. An example of each window is shown in Table 4.1.

Table 4.1 Example of different region

	Bag-of-words	Father word in dependency tree	String of words in dependency tree
America and Iraq	和(and)，英国(Britain)，两国(two countries)，联军(allied forces)，出兵(attack)	出兵(attack)	联军(allied forces)，出兵(attack)
America and British	和(and)	联军(allied forces)	联军(allied forces)，和(and)
British and Iraq	两国(two countries)，联军(allied forces)，出兵(attack)	出兵(attack)	和(and)，联军(allied forces)，出兵(attack)

4.1.3 Sentiment Recognition

In this step, we seek for the sentiment polarity between entities based on the relative region. The task can be seen as predicting higher level sentiment from lower level (word). It is not an easy task as researchers had reported that higher level sentiment is not necessarily consistent with lower level sentiment [58-60]. Some words such as “extremely, but” could change or even reverse the sentiment of higher level. Our previous research had met the problem as well. We had tried rule based method for sentiment recognition and sentiment of father word that attaches two countries decides sentiment of the country

pair. It got precision of 79.66% in two-category solution. This simple method did not perform well especially when dealing with negative examples. As the effects of privative words are ignored, the method only gets 64.52% precision in negative examples. Expanding the region to string of words and considering effect of privative can improve the performance. For example, some teams of COAE2011 used rule based method on string of words to deal with the problem. If there are privative words in the string, the total polarity of sentiment will be reversed. It obtained better performance than the method that did not consider privative words.

As sentiment of language has the characteristic of complicated and nonlinear, especially in Chinese. Rules cannot cover all conditions. Machine learning approaches based on statistics perform well in the field [61]. It learns information from training data and could be robust for noise. The information learned from labeled data instead of rigid rules is used while dealing with new data. There are several methods of machine learning for classification such as support vector machine (SVM), Decision tree etc. Our task is different from traditional linear classification problem. Different word plays different role in sentence and has different effect on sentiment. Their interactions also affect final sentiment. Privative that reverses whole sentiment of sentence is a typical example. We use a probabilistic model based on graph to solve the problem. In the graph each word is a node with features and relations are represented by connections between nodes. Both nodes and connections affect final result. A lot of work had been published in this model and different machine learning approaches based on the model are proposed as well. We choose conditional random fields (CRF) in our research for it has the advantages of non-independent features and solves the problem of label bias. Details of CRF will not be described in this paper as there are enough papers and mature tools about CRF [62]. In our research we focus more on the effect of machine learning method and features that affect

sentiment. The CRF++－0.50 free tool based on LBFGS training method is used for our research[63].

Before using CRF++, we have to specify the feature template. It describes features used in training and testing. For a word in the sentence after parsing, it has attributes of part of speech (POS) and syntactic component. The POS plays an important role in sentiment as we mentioned before, most sentiment words are verbs, nouns and adjectives. Syntactic component reveals relations between words and its effect on sentiment should be considered as well. In addition, different sentiment function of word may also affect result. We set four types of words of different functions which are ordinary, sentiment, modifier and privative. Sentiment words are words in our emotional word dictionary that express sentiment. Modifier words are words that can strengthen or weaken the sentiment. Privative words can reverse the whole sentiment. At last, different values of sentiment words that represent intensity of sentiment maybe another factor worth consideration.

Finally, five features of word itself Wi, POS Pi, syntactic component Ci, sentiment function Fi and value of sentiment Vi are written in training template. Relations between words also exist in the template as shown in Table 4.2.

Table 4.2 Feature template

1	Wi
2	Wi-1
3	Wi+1
4	Wi-1& Wi
5	Wi& Wi+1
6	Wi& Pi&Ci&Fi& Vi

One thing should be mentioned is that the i-1 -th word is the one that the i-th word relies on in dependency tree, not the previous word of the i-th word in the sentence. We can get a CRF model using CRF++ tool based on feature template and training corpus. After obtaining the model, sentiment of relations could be assigned automatically. The algorithm is shown in Fig 4-1.

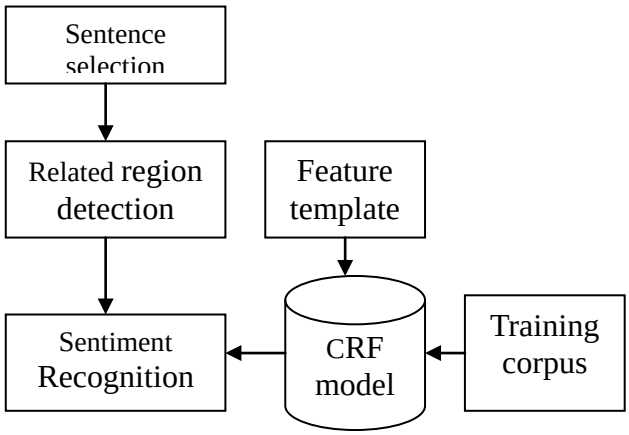


Fig 4-1: Overall framework

4.1.4 Related Resource

In proposed research, two emotion dictionaries are used. One is the emotion dictionary mentioned in Chapter 3, the other a modifier and privative dictionary used to identify the modifier and privative words. Different from emotion words, the number of modifier and privative words is not large, so we obtain these words come from the collection of COAE2011 with some additional words from Internet [43]. There are 85 modifier words and 24 privative words in all.

4.1.5 Experiments and Result Analysis

There are no ready corpus for recognizing sentiment between entities at present. We still use important international current affairs of "China Comment" as corpus for the experiment. We collect news of three years (2009 to 2011). Sentences containing more than one country are selected in the news, and then we extract country pairs from these sentences. A total of 766 pairs are extracted and analyzed. With the help of two human annotators, 290 pairs are labeled with positive sentiment, 184 pairs are labeled with negative sentiment and the rest are neutral, as shown in Table 4.3.

Table 4.3 Statistical information of country pairs.

Sentiment	Number of country pairs	Percentage
positive	290	37.9%
negative	184	24.0%
neutral	292	38.1%

In these pairs, the coordinate pairs are abandoned as mentioned in section 3.4.2. We show some pairs of different sentiment from our data corpus in Table 4.4.They include all three types of sentiment which are positive, negative and neutral.

Table 4.4 Some pairs of different sentiment

Pair and Sentiment	Sentence
<Thailand Palestine> positive	泰国外交部近日宣布，泰国已经承认巴勒斯坦为一个独立国家 The foreign ministry of Thailand declared that Thailand had recognized Palestine as a sovereign nation
<South Korea North Korea> negative	韩国当天发表政府声明对此进行谴责，并称朝鲜的行为是坚决不能容忍的挑衅行为 South Korea condemned it through a government statement and said the action of North Korea is an unacceptable provocation.
<Taiwan Hong Kong> neutral <Taiwan China> neutral <Hong Kong China> neutral	台湾地区共有 24 位上榜，香港有 37 位，中国大陆则有 95 位亿万富豪上榜 There are 24 billionaires from Taiwan, 37 from Hong Kong and 95 from mainland China in the ranking.

In order to compare the effect of machine learning, we design four methods with different related regions and principles. All these methods recognize the sentiment of country pairs and give a final value of -1,0 or1 that represent negative, neutral and positive.

BW-R Method: bag-of-words region and rule-based

The related region is the set of words in the sentence between two countries of a pair. The sum of sentiment value of each word in region decides final sentiment polarity. Sentiment value of word can be found from emotional dictionary.

WD-R Method: word in dependency tree region and rule-based

The related region is the father word of two countries in dependency tree. Sentiment of the father word is the sentiment of the country pair.

SD-R Method: string of words in dependency tree region and rule-based

The related region is the string of words that attaches two countries in dependency tree. We sum sentiment value of each word in the string to get the final sentiment polarity.

SD-C Method: string of words in dependency tree region and CRF-based

The related region is the string of words that attaches two countries in dependency tree. Then words in the string are transformed into the form of CRF template with features of word itself, POS, syntactic component, sentiment function and value of sentiment. We split corpus into five portions, four are used for training and one for testing. CRF model after training is used to annotate sentiment polarity of testing pairs.

The accuracy of different method is shown in Table 4.5.

Table 4.5 The accuracy of different method.

Method	Accuracy
BW-R Method	0.335
WD-R Method	0.454
SD-R Method	0.428
SD-C Method	0.835

The accuracy of SD-C method is calculated as the number of correct pairs annotated in testing corpus divided by the number of testing pairs.

The same testing corpus is used for experiments of other methods as well. SD-C method performed well, especially when dealing with long sentence of complicated interaction. We analysis an example to show why SD-C method can perform well, the sentence is “据驻俄罗斯大使馆介绍，截至目前没有中国人在莫斯科地铁爆炸事件中伤亡的消息. (Based on embassy in the Russia, no news said people from China are injured or dead in the explosion of Moscow subway up to now)”.From the sentence we get that there are no injuries or deaths, so the sentiment should be positive. However, as the sentence is stated with privative and sentiment of words in related region shows negative. BW-R, WD-R and SD-R methods cannot obtain correct final sentiment. SD-C methods can consider the effect of privative “没

有(no)” correctly. Therefore, it can recognize the sentiment correctly. The related region of each method and final sentiment is in Table 4.6.

Table 4.6 Analysis of different method.

method	Region and sentiment	Final sentiment
BW-R Method	大使馆（embassy）=0.0 介绍(introduce)=0.0 截至(up to)=0.0 目前(now)=0.0 没有(no)=-0.0527	negative
WD-R Method	消息(news)=0.0	neutral
SD-R Method	人(person)=0.0 没有(no)=-0.0527 伤亡(injuries or deaths)=-0.15 的(about)=0.0 消息(news)=0.0	negative
SD-C Method	人(person)=0.0 没有(no)=-0.0527 伤亡(injuries or deaths)=-0.15 的(about)=0.0 消息(news)=0.0	positive

In previous work of two-category solution, WD-R method got accuracy of 79.66%. However, in this task of recognizing sentiment into three types, WD-R method did not have satisfied result and only got accuracy of 0.454. Sentiment of neutral can be separated into both positive and negative in two-category solution. That is why WD-R method has much higher accuracy in previous work. From Table 4.5, we can see that machine learning method obtain the best performance. Comparing with other methods, it improves accuracy about 40% percent which is a striking result. Different related region affects result as well. String of words in dependency tree shows best performance and bag-of-word is the worst.

We analyze errors of each method as well. Percentage of each sentiment in errors is shown in Table 4.7. Different method shows different error composition. For example, in BW-R method half of errors are neutral. However in SD-C method, three types of sentiment occupy nearly equal percentage of errors.

Table 4.7 Analysis of errors

	BW-R Method	WD-R Method	SD-R Method	SD-C Method
positive	11.3%	32.1%	16.3%	38.0%
negative	35.1%	36.8%	30.8%	29.7%
neutral	53.6%	31.1%	52.9%	32.3%

Error rate of neutral is extremely high in BW-R method and SD-R method. In these two methods, final sentiments are dependent on sum of related words. Even if there is only one word in related words showing sentiment, the final result will not be neutral. Therefore, the two methods cannot recognize neutral sentiment well. We mentioned before that WD-R method did not perform well especially when dealing with negative examples as a result of ignoring effect of privative words. Comparing WD-R method and SD-C method, we can see that error rate of negative is reduced obviously in SD-C method. It reflects that CRF method could consider function and interaction of words. It can improve the performance of recognizing negative sentiment indeed.

4.1.6 Parameters of CRF

From the above result, it is suggested that SD-C method based on CRF obtain the best performance. Different parameters, such as numbers of training corpus and features used in training will affect performance of CRF [64]. We carry further experiments on CRF to discuss effect of different parameters.

Firstly, we analyze the effect of training corpus. In all 766 pairs, 152 pairs are used for testing and the other 614 pairs are in training corpus. We select different number of pairs in training corpus for training and obtain the result in Table 4.8.

Table 4.8 Different number of training pairs

Number of training pairs	Accuracy
5	0.331
10	0.392
50	0.471
110	0.615
150	0.652
200	0.728
300	0.781

400	0.821
500	0.830
600	0.835
614	0.835

With the number of training pair increasing, the accuracy is increasing as well. However, the rate of increasing becomes slower and slower. There is almost no change of the accuracy when number of training pairs expands to over 400. The result also proves that we have trained enough instances in CRF model to cover the testing. The scale of training corpus is proper.

In SD-C Method, we choose five features in training template which are word itself, POS, syntactic component, sentiment function and value of sentiment. In order to evaluate effect of different features and find the best features, we write different training templates and test them on the same corpus.

The first experiment changes features of row 6 in Table 4.2, five features are reduced to two which are word itself and one other feature. We also evaluate performance of single feature of word itself. The result is shown in Table 4.9.

Table 4.9 Effect of different features (1)

Features	W	W+P	W+C	W+F	W+V
Accuracy	0.812	0.845	0.846	0.857	0.852

From Table 4.9, we can see that adding one feature can improve result and the performance order of each features is F>V>C>P. In other words, sentiment function is the most important feature. In order to verify the conclusion, other experiments are implemented. We also change features of row 6 in Table 4.2 and delete one feature in each experiment. Four features are under consideration in these experiments. Results are shown in Table 4.10.

Table 4.10 Effect of different features (2)

Features	All-P	All-C	All-F	All-V
Accuracy	0.866	0.844	0.828	0.841

Deleting sentiment function affects performance most. Based on Table 4.9, the performance order of each features is F>V>C>P as well. This result is not

astonishment.POS and syntactic component are general syntactic features while sentiment function and value of sentiment are special features of sentiment. Value of sentiment decides sentiment of each word which is the base of predicting higher level sentiment. Sentiment function considers function of different word on higher level, such as reversing and weakening. It can deal with privative word that reverses the whole sentiment in higher level, and improves performance of our experiment significantly. We also note that most results in Table 4.9 and 4.10 are better than the result of five features in Table 4.5. It indicates that too many features may reduce the performance to some extent. In order to find the best features, we also test the combination of three features. W+F+V obtain the best result with the accuracy of 0.868.

4.2 Collect Data of News on Internet

We constructed an international network from a relative small corpus that is news from magazine in prototype system. The scale of corpus is still not big enough. As we know, Internet is the biggest information source in the age [65, 66]. Countless number of news about international situation emerge on Internet every day. We tried to collect news on Internet as data source in the improved system.

As number of news on Internet is so huge that cannot be collected by human, we collect with the help of baidu news search engine [67]. As the biggest search engine in China, baidu provides special search of news. Its interface is shown in Fig 4-2. Users can set variables of keywords, time periods, keywords in title or in text, way of sorting, number of result and so on.

Fig 4-2: Interface of baidu news search

The query sentence is as following if we search “中国” in title of news and set time from 2013-1-1 to 2013-2-1.

http://news.baidu.com/ns?from=news&cl=2&bt=1356969600&y0=2013&m0=1&d0=1&y1=2013&m1=2&d1=1&et=1359734399&q1=%D6%D0%B9%FA&submit=%B0%D9%B6%C8%D2%BB%CF%C2&q3=&q4=&mt=0&lm=&s=2&begin_date=2013-1-1&end_date=2013-2-1&tn=newstitledy&ct=0&rn=20&q6=

Fig 4-3: Result of search

Fig 4-3 is the result of the search. We can find that in result, it can provide not only title of news but also number of hits retrieved by the query. Based on the baidu news search engine, we define two functions.

BaiduNewsHit: It provides the number of news hits that retrieved by given keywords.

BaiduNewsTitle: It provides k titles of news that retrieved by given keywords.

These two functions play crucial roles in our search. Following steps are on the basis of the two functions. There are over two hundred countries in the world. In order to construct the international network, we should collect data about each countries and each in every two countries. From a list that contains all countries and the United Nations (UN), we set names of each country as keyword to obtain BaiduNewsHit and BaiduNewsTitle. For country pairs, we set names of the two countries as keyword to obtain their BaiduNewsHit and BaiduNewsTitle as well. After all, we obtain data of 233 countries and more than 50000 country pairs. A few missing data is due to no response of query. In order to study international network in different period of time, we also collect data that set time in September and October of 2012. Through calculation of these data, we can construct the international network.

4.3 Improved System

With two main improve points discussed above, we propose our improving approach of prototype system. The approach is still based on the five types of international relation which are defined in Chapter 3. The classification of them are based on three attributes which are intensity, quality and status of relations as well. However, news collected on Internet has some unique features, improved approaches are proposed to fit the data of news collected on Internet.

Intensity :

The approaches of estimating the intenstiy of relation are generally based on co-occurrence of two countries. We mentioned before that several indices can be used to measure the co-occurrence. As we only obtain occurrence of each countries and co-occurrence of every two countries, a special index called overlap proportion. It is defined as following:

Definition: Overlap proportion
Nc1←BaiduNewsHit <“C1”>
Nc2←BaiduNewsHit <“C2”>
Nc1∧c2←BaiduNewsHit <“C1 C2”>
Overlap proportion←Nc1∧c2/ Nc1+ Nc1∧c2/ Nc2

Fig 4-4: Get intensity of relations

Overlap proportion can reflect intensity of relation relatively fair. In some case, low co-occurrence (BaiduNewsHit <“C_1 C_2”>) did not imply weak relation; it is due to low occurrence (BaiduNewsHit <“C_1”>). Overlap proportion still get high value in this situation if proportion of co-occurrence is high.

Quality:

Quality indicates relations are friendly or conflict. We use sentiment analysis to detect the quality of relations. The method is based on title of news. Through a title containing countries, we should evaluate if the relations among them are positive or negative. In our work, we had tried different methods of recognizing sentiment of relations among entities and compared their results. The details of them can be referred in Chapter 4.1. A CRF-based method using string of words in dependency tree as related region gains the best performance. We use this method with some simplification. Top ten titles of each country pair are selected first. Then, a value of -1,0,1 is assigned to each title through the method. The value reflects negative, neutral and positive relation. Total ten values indicate quality of relation between two countries.

Comparing with method mentioned in Section 4.1, we did not need to select sentence that contains countries (Titles for analysis already have two countries). Hence, the general algorithm of assigning values to titles can be divided into two steps. First, we use parser tool of LTP to analyze title of news and obtain related words of two countries. Some syntactic attributes of words are obtained as well, such as part of speech (POS) and syntactic component. In the second step, a well-trained CRF model is used to assign values based on five features. Five features are word itself W_i, POS P_i, syntactic component C_i, sentiment function F_i and value of sentiment V_i. The feature template is as shown in table. Modifier and privative words that used to decide sentiment function come from the collection of COAE2011 with some additional words from Internet. The value of sentiment is based on a large emotion annotation corpus (Ren-CECps).

Status:

Status indicates the balance of relations between two countries. If a bilateral relation plays an important role in diplomacy of one country but not so important to the other one, we assert that their status is unbalance with the feature of subordination. For example, country C1 and C2 contact each other five times in one year. Country C1 has a total foreign intercommunication of ten times in the year while country C2 has one hundred times. Obviously, their relation has the feature of subordination and the bilateral relation is more important to country C1. We use subordination coefficient to evaluate status.

Definition: Subordination coefficient
Nc1←BaiduNewsHit <“C1”>
Nc2←BaiduNewsHit <“C2”>
Nc1∧c2←BaiduNewsHit <“C1 C2”>
Subordination coefficient←Nc1∧c2/ Nc1-Nc1∧c2/ Nc2

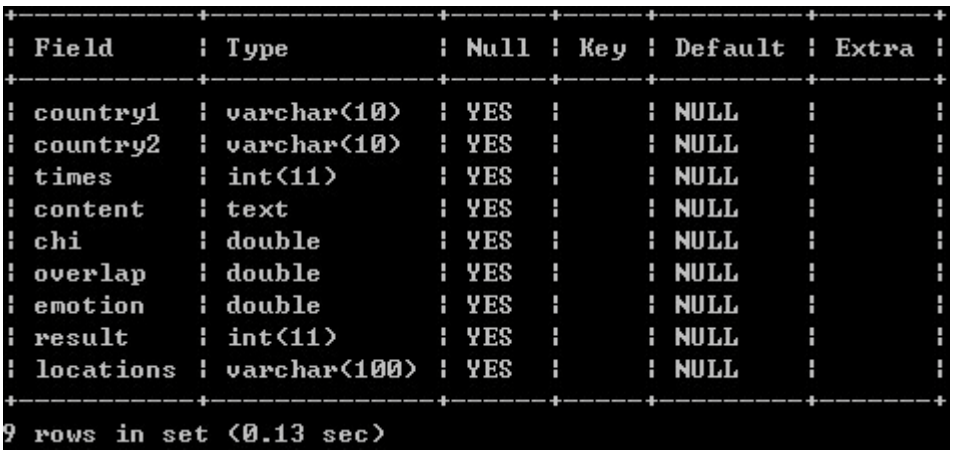
Fig 4-5: Get status of relations

An international network is obtained while getting nodes and connections. Nodes are

233 countries. We did not assign connections to each pair of countries as there will be more than 50000 connections. Obviously, not all pairs have enough interactions. Only top 400 pairs with high co-occurrence are assigned a connection. We will explain it is reasonable and covers most interactions in the world.

4.4 Data and Analysis of Improved System

In order to study international networks in different time period, we collect data with three different times setting which are no limitation, September in 2012 and October in 2012. They are stored in Mysql database.



Field	Type	Null	Key	Default	Extra
country1	varchar(10)	YES		NULL	
country2	varchar(10)	YES		NULL	
times	int(11)	YES		NULL	
content	text	YES		NULL	
chi	double	YES		NULL	
overlap	double	YES		NULL	
emotion	double	YES		NULL	
result	int(11)	YES		NULL	
locations	varchar(100)	YES		NULL	

9 rows in set (0.13 sec)

Fig 4-6: Structure of database

The structure of database is in Fig 4-6. It contains all the information, such as names of countries, time of hits and news titles of hits. Three features of relations and final types of relations are contained in the database as well. Through searching and calculation of the database, we can analyze data easily. The data about BaiduNewsHit of single country are in Table 4.11.

Table 4.11 BaiduNewsHit of single country

No time limitation			September 2012			October 2012		
Total hits: 63935698			Total hits: 777097			Total hits: 1218232		
hits	Number of countries	Total hits	hits	Number of countries	Total hits	hits	Number of countries	Total hits
>102	216	63935153	>10	180	776988	>10	180	1218114

>103	192	63923000	>102	139	774946	>102	142	1216427
>104	151	63718900	>103	62	744310	>103	66	1188280
>105	64	60356000	>104	15	565000	>104	13	989500
>106	12	44280000	>105	2	258000	>105	1	617000

From the data, we can see that although hits of three time periods are different, the distribution of them are similar. Most countries have average hits and more than 100 countries are in the middle two levels. We also find that a few countries obtain most hits. About 25% of countries obtain more than 90% of hits; it indicates that focusing on only a few countries can cover most events in the world.

Table 4.12 BaiduNewsHit of country pair

No time limitation			September 2012			October 2012		
Total hits: 14890288			Total hits: 212915			Total hits: 193755		
hits	Number of pairs	Total hits	hits	Number of pairs	Total hits	hits	Number of pairs	Total hits
>102	5738	14603325	>101	1239	207592	>101	1096	188499
>5*102	2493	13846338	>5*101	514	190064	>5*101	472	173322
>103	1589	13192990	>102	315	175265	>102	293	160321
>5*103	453	10767640	>5*102	77	125869	>5*102	72	110494
>104	233	9219500	>103	32	93250	>103	26	79860
>5*104	40	5077900	>5*103	4	49870	>5*103	4	38960
=0	35014	null	=0	51274	null	=0		

Comparing with hits of single country, hits of pairs have different characteristic. Hits are more concentrated; more than 80% hits belong to about 1% pairs. However, most pairs have low hits or even zero. It indicates that the international network is sparse and connections only exist in a few country pairs. Therefore, we set a threshold of top 400 pairs. For the top 400 country pairs with high hits we assume they have relations and assign connections to them. Comparing with threshold of hits, threshold of fixed pairs can ensure the scale of international network in different time period. The threshold of 400 can cover most events happened between two countries in the world. Through calculation of 400 pairs, we find that the 400 pairs include more than 60 different countries in all three time periods. It shows that this threshold can cover enough hot countries as well. For the 400 pairs, we analyze their overlap proportion, sentiment analysis, subordination coefficient and final

types of relation.

Fig 4-7 shows distribution of overlap proportion. Most values (over 200 pairs) concentrated from 0.1 to 0.01. Comparing with time period of one month, no time limited has less extreme value (>0.2 or <0.01). High values often exist between neighbors, such as <Palestine, Israel> and <South Korea, North Korea>. Powerful countries and minor countries usually have high overlap proportion as well. For Australia and Nauru, nearly all international communications of Nauru are with Australia and high overlap proportion is obtained. As we expected, low values exist between powerful countries without much communications, such as <South Korea, Russia> and <India, Germany>.

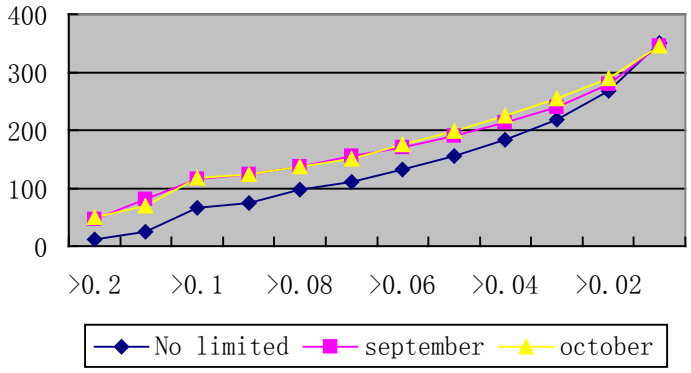


Fig 4-7: Distribution of overlap proportion

Distribution of subordination coefficient is similar to that of overlap proportion although it is more scattered. As in Table 4.13, most values are located between 0.1 and 0.005. Without doubt, high values exist between powerful countries and minor countries such as <Russia, Georgia>, <America, Djibouti>. We also find lots of high values are about China as news are obtained through Chinese searching engine, it focus more about other countries' commutations with China.

Table 4.13 Distribution of subordination coefficient

No time limitation		September 2012		October 2012	
|value|	Number of pairs	|value|	Number of pairs	|value|	Number of pairs
>0.2	9	>0.2	34	>0.2	36
>0.1	52	>0.1	92	>0.1	100

>0.08	81	>0.08	115	>0.08	118
>0.05	126	>0.05	153	>0.05	163
>0.03	171	>0.03	193	>0.03	213
>0.01	260	>0.01	277	>0.01	292
>0.005	315	>0.005	323	>0.005	330

Sentiment value is obtained through 20 pieces of news title about country pair. So the range of value is from -20 to 20. Analysis of sentiment value is in Table 4.14.

Table 4.14 Distribution of sentiment value

No time limitation		September 2012		October 2012	
value	Number of pairs	value	Number of pairs	value	Number of pairs
>-15	395	>-15	395	>-15	399
>-10	386	>-10	390	>-10	395
>-5	351	>-5	332	>-5	374
>0	222	>0	184	>0	190
>5	107	>5	49	>5	38
>10	37	>10	14	>10	3
>15	7	>15	1	>15	1
>0	222	>0	184	>0	190
<0	115	<0	105	<0	106
=0	63	=0	109	=0	103

We can find that most values located in the middle level (-5 to 5). Positive values are more than negative values. It indicates the general international relation is peace. Country pairs with negative values are corresponding to common sense, such as <America, Sudan>, <Lebanon, Israel>.At last, we classified relations into five types based on Table 4.15.

Table 4.15 Classification of relations

Relation	Overlap proportion	Sentiment value	Subordination coefficient
Cooperation	≥0.02	≥3	|value|≤0.1
Subordination	≥0.02	≥3	|value|>0.1
Conflict	≥0.02	≤-3	*
Confrontation	≤0.01	*	*
Neutral	0.01<value<0.02	*	*
	≥0.02	-3<value<3	*

Numbers of each relation are in Table 4.16. As discussed above, most values of three features are in the middle level. It is not amazing that number of neutral is the most. Numbers of other types are generally average. However, networks of different relation has different features, we will discuss it in next chapter.

Table 4.16 Numbers of each relation

No time limitation	September 2012	October 2012

relation	Number of pairs	relation	Number of pairs	relation	Number of pairs
Confrontation	54	Confrontation	56	Confrontation	62
Neutral	172	Neutral	229	Neutral	223
Conflict	52	Conflict	32	Conflict	33
Subordination	33	Subordination	38	Subordination	32
Cooperation	89	Cooperation	45	Cooperation	50

Chapter 5 Visual interface

5.1 Introduction

A visual interface is an essential part of our work as well. The interface can exhibit international network in the form of graph with features of clear and vivid. Similar to e-diplomacy Hub, we also draw the interface on the world map as it can refelct the real location of countries and easier to grasp hotspot.

In this chapter, we are going to show the international relation network constructed in previous chapters in a visualized way. Similar to e-diplomacy Hub, the interface will exhibit international network in the form of graph on the world map. As an indispensible step in international network research, visualization has two significances. One is to exhibit the constructed international relation network and its various relation features in a visual interface directly. The other is to take advantage of human's imaging thinking, so as to improve people's cognitive ability, help researchers to unveil rules and features of international relations.

From a psychological prospect, visualization is a mental process to recognize objective things via visual organs and then form images in one's mind. Compared with other recognition methods, visualization improves people's observation ability and helps to form concepts on the whole. This provides great convenience for comprehension and memory and leads to an incomparable advantage over other methods. Visualization technology is developed from this psychological base. It is a technology to transfer data to images, show them on a screen and interactive process them with the help of computer graphics and image process techniques [68, 69].

Work on visualization with computers origins in the middle of 20th century when researchers created a pile of graphs to illustrate data process results with early computers. Research on visualization thrives in 1980s with the increase of computer capacity. In 1987, a seminar on visualization was held by National Science Foundation of the United States. Research on visualization was focused on scientific and engineering computing problems at that moment. Since the scale of data produced by computers has reached a level people could hardly imagine, it is an urgent need to show these data in static or dynamic ways with the help of computer graphic information. This provide convenience to users to recognize the data from a overall and direct view, thus to understand the data correctly and timely, and finally to lead users to correct results [70]. IEEE has been holding conferences on visualization each year since 1990. According to incomplete statistics，there are over 50 research groups working on data visualization , more than 10 special journals and magazines and a great number of international conferences or forums all over the world.

We are going to directly exhibit international relation networks with information visualization technology, as well as to assist knowledge finding in text mining with visual interface and interactive methods. Physiological researches indicate that more than half neurons in human brains are visual-related, and more than 60% external information is obtained through eyes. Humans are born with effective and high-capacity image processing channels which can perceive, recognize and analyze objective environment from their visions. Experiences show that human visual system is far more capable to perceive and understand images quickly than to recognize and process texts. Due to the limits of these experiments, it is not clear how the graphic information enters people’s brains and how the brains react. However, there is no doubt that direct images have great edge over texts and signals on cognitive matters. In summary, visualization technology has the following advantages:

Vivid, intuitive and visible: visualization technology exhibits data to users in images, graphs or animation models in a vivid, intuitive and visible way. This provides users a foundation to search and analyze information directly.

Multi-dimensional data integration: different features and variables can be integrated in one database and analyzed from multi-dimensions.

Interaction ability: visualization offers an environment where data can be managed and processed flexibly and intuitively. Users can adjust variables flexibly, understand information in different point of views, or make data compares and analysis, with the help of which they draw their conclusions or make their decisions.

Due to the above advantages, visualization technology works better when expressing complex structures such as networks. Although we can store networks in computers in data graphs like matrixes or texts, it is for human beings quite hard to understand these data forms. “A picture is worth a thousand words” —visualization has great advantage to exhibit networks. Researches on social networks and complex networks indicates that visualization technology is able to present massive complex network information to users effectively and directly. It can also turn boring data into acceptable and understandable forms. It provides great convenience to information search and inquiry by presenting the whole network structure accurately. What’s more important, it unveils valuable rules and knowledge hidden behind which cannot be discovered by traditional methods.

An innovational international network visualization method, which is built on relative network visualization research results, international network features and electronic maps, will be introduced and realized in this chapter. In order to create favorable conditions to discover new knowledge, we will also add some interactive functions such as searching in the visualized interface. The rest of this chapter is as follows. Part two introduces recent research achievements on network visualization. Then we move on to our method to build

the visualized interface in part three, including system function interface design and steps to realize it. In the last part, an international relation network gained before will be shown in our visualized interface, through which we will analyze several features of the network structure.

5.2 Related Work

Network visualization is an important field of visualization technology research. Due to the complex structure and contents of network structures, it would be hard to reflect the information contained in network structures by tables or texts [71, 72]. Researchers therefore attempt to exhibit network data in images and graphs, taking advantage of human being's visual perception, to help people recognize network structures and discovery valuable information. More and more attentions have been paid to network visualization works since 1990s, and several theoretical and practical results have been achieved.

Research on network visualization involves aesthetics, graph theory, statistics, information visualization, human-computer interaction and so on [73, 74]. Visualization algorithm is paid the most attention among those research theories now. It is about how to organize the structure of networks and positions or shapes of spots and lines, so that users could get the best perception effects. Practice shows the layout of networks is significant to users in understanding the networks structures. If the layout is not designed properly, even a small network consisting of less than one hundred nodes would become too complex to understand. Since the early ages of network visualization, the best design to draw networks, from which users could get the best perception effects to understand information in the networks, has always been pursued by researchers.

Practical tools on visualization boost in recent years, including all kinds of visualization frameworks and class library, user developed special network visualization tools, general

network visualization tools and so on.

General network visualization tools are the most popular network visualization tools. Usually, they are developed by professional companies, with abundant functions and great practicability. Instead of limited to certain tasks, they are able to deal with networks with general features. They also provide functions such as software integration, interactive processing, classification statistics, focus analysis and so on. Users can analyze and research the network from different dimensions or different points of views. Some famous general network visualization tools are Graphviz, Piccolo, UCINET and Toolkit.

Take UCINET as an example [75]. UCINET is a fully-functional network analysis integration software which contains functions such as network analysis, visualized exhibition and so on. It reads network source data in matrix form, conducts analysis such as centrality analysis and cluster analysis, then outputs the results to the drawing module and finally visualize it by software like Net Draw. HÄMMERLI used UCINET 6 to analyze and visualize the relationship network of the Chechnya War. The conflict relationship network is visualized as Fig 5-1.

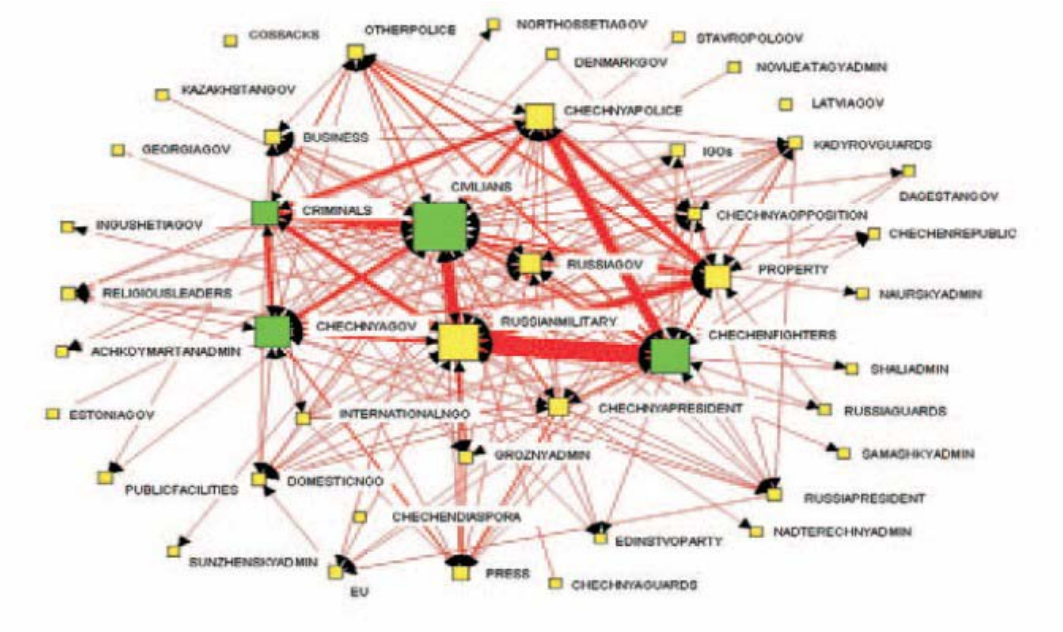


Fig 5-1: An example on relationship network visualization [21]

As a branch of social network research, international relationship network research is now doing the same in visualization. Usually, researchers perform the whole process from data analysis to visualization with the help of general visualization tools—they use spots to stand national entities, lines to relationships, and realize international relationship network visualization, ignoring the geographic information of states. With the help of mature social network analysis tools, this seems to be a shot cut. However, a vital point of international relationship network is that states do have geographic locations, which means unlike usual social networks, geographic locations have significant influences to international relationships. Here we take an example of international trade network Fig 2-1. Geographic locations and the distance between two countries plays an important role in trade network. Direct trade meets much more difficulty if traders have to travel far. That is why some countries at crossroads are born trading centers. Unfortunately, you can barely see this in this trading network, just because the geographic information is omitted. We hope to overcome this shortcoming by fully considering geographic information when visualizing international relationship networking. We will combine international relationship network and map geographic information to full disclose network information. Taking advantages of maps' spatial recognition and information assistance ability, users' recognition process will be simplified and thus they will be quicker to recognize international relationships.

Although it is not common to realize international network visualization with geographic information in network visualization researches, researches realize data visualization and interactive process are not rare in Geographic Information System (GIS), which leaves us affluent achievements.

Map visualization is based on relative recognition theories, too. It admits that people gain information more effectively and efficiently from visualized images than from texts, symbols or data. Someone believe that one page of map contains information as a 20 pages

book of the same size does. What's more, human's visual ability is the most powerful tool to filter information and scan important patterns. People could easily form an overall recognition about a picture by combining features such as shape, color, brightness, movement, vector, quality and so on. Taking advantage of this, maps translate different information from the real world into visual grammars, and express information vividly by graphic metaphor. That's why people are able to get to know the world quickly and comprehensively from a single map. To match the abstract symbol expression with human recognition and pass the information as much as possible, researchers, from the prospect of psychology and recognition points out the following matters which should be paid attention to when visualizing maps[76-79].

1. Quantity of symbols in a map. Since human beings' ability to react to symbols is limited, one may be blunt to symbols or even omit them if too many symbols are given at the same time. Thus, when designing visualized maps, people's capacity to react to symbols should be considered, in case that the map carries too many symbols and too much information.

2. Differences between symbols and their features. Maps differ information of the real world by different symbols and their features. The smaller the difference among symbols is, the longer it takes to react to these symbols. As a result, you must fully consider the difference among data, chose the suitable symbols and make a difference, so that users can grasp important information as soon as possible.

3. Compatibility of symbols. Information stood by symbols should agree with people's daily customs or expectations. This would make users to react faster to these symbols for they cope with people's common sense. For example, we often use red to stand for danger, so urgent or serve situations should be in red in maps, and peaceful situations can be mapped in green.

4. Consistency. It is an important issue that people react fast at the begin of the activity, but get slower and slower afterward. This means that dynamic activities should not last too long.

Interaction is another problem referred many times in nowadays' map visualization research. Compared with traditional maps, visualized maps pay more attention to the 'human-map' relationship, which is people-oriented. The maps are supposed to react to users' behaviors and offer multi-dimensional views on different geographic information. The main purpose of this interactive ability is to provide an environment to analysis and discover. Currently available interactive functions for map visualization are assignment formulation, joint list, highlight, zoom, fisheye focus, time series and so on. These interactive functions provide conditions for discovering knowledge in the next step. Users can then set different parameters and variables with the interactive functions, compare or connect different parts of the information, test their analysis and finally find new rules or knowledge. A series of new knowledge can be found based on maps and relative geographic information visualization, including statistic features of the target(i.e. quantity and size), spatial distribution of the target, association rules among the targets(i.e. connection, inclusion, symbiosis), evolution over time and so on.

5.3 Visualization System Design

5.3.1 Functional Design

The purpose of the visualization system in our work is to exhibit the international network that constructed before. The visualization system can help users check and use international network. Plaisant summarizes the specific tasks of network visualization as followed [80]:

1. Topology structure of network: help users to see the total structure of network

2 . Node attribute: help users to check attribute of specific node, search nodes with

specific attribute

3 . Network browsing: trace special path or node

4. Network overview: help users acquire the global information of network

However, not all network visualization should have such functions. The main purpose of our system is to show the international network with their geographic position. It also should provide interactive function to help users check network from different aspects. Therefore, we design such functions in our visualization system.

1. Showing the total network with world map
2. Checking relation between two countries
3. Checking relation of one or two countries
4. Searching network of special relation
5. Image control, zooming the interface

5.3.2 UI Design

Based on the function, we design the UI. It separated into two parts: control region and exhibition region. Control region is in the upper left corner. Users can select different ways of observing the network from different aspects. Exhibition region is in the middle to show countries and their relations.

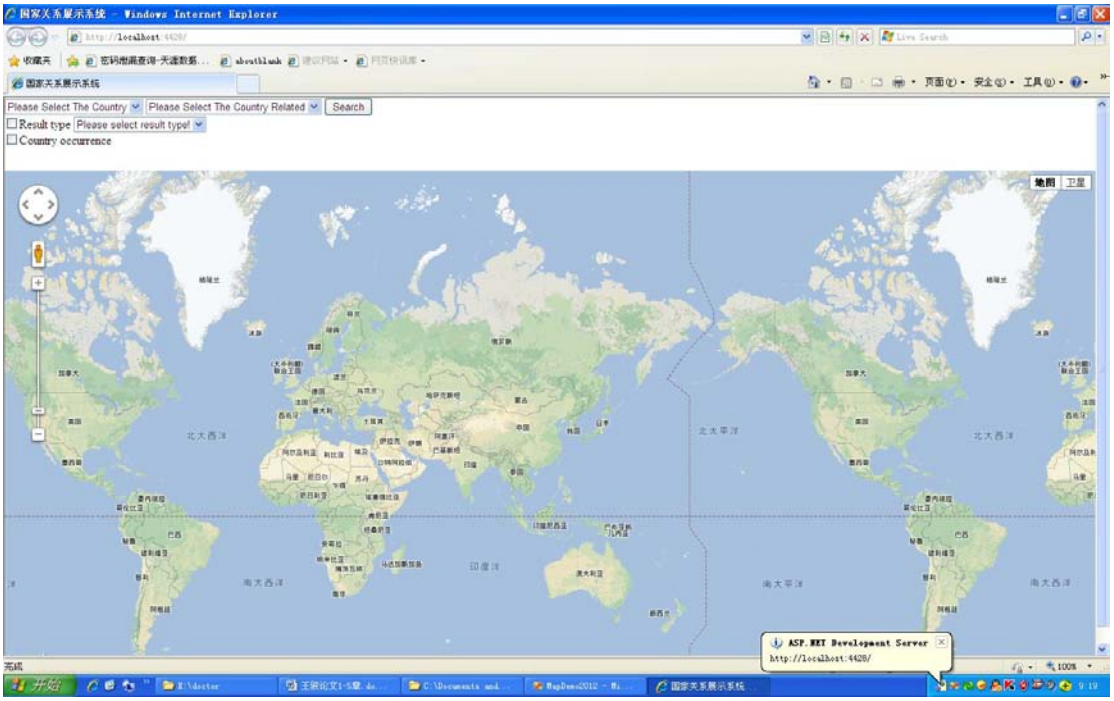


Fig 5-2: UI design

5.3.3 Architecture Design

We build our visualization system based on WebGIS. WebGIS are called web mapping sometimes. It combines Internet with mapping process. Through WebGIS, users can post map on line, create visualizations and display data to public [81,82]. Generally, WebGIS composed of four parts: client, database, web service and GIS application service.

Client: the interface between user and system. It responses actions of users, submits request to the web service

Database: the warehouse of data

Web service: the middle lay. It transmits request form client to GIS application service and returns results.

GIS application service: processing center of the system. It processes the requests and gets results.

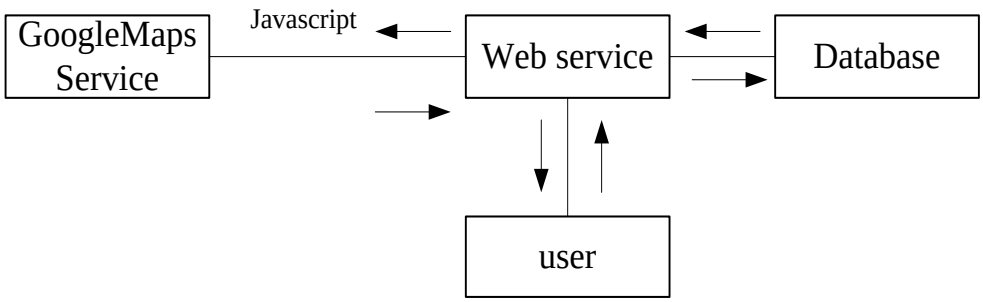


Fig 5-3: Architecture of visual system

We survey several architectures of WebGIS and choose B/S architecture to realize the four parts in our system [83-86]. Client is the browser of user. Database is based on mysql database. In order to improve the speed of response, we design an individual database. The database contains fields of country pairs, intensity, sentiment, subordination, type of relation and locations of countries. IIS web service that integrated in windows is the web service in our system. For GIS application service, we choose Google Maps API. Google Maps API allows users to integrate Google map to their own web, the API also provides relative function such as zooming, adding symbols [87, 88].

The general work framework of visual interface can be summarized as followed. While the program is started, Google map is loaded to user's browser first. Then users can select different function of the interface to search different aspect of international network. User's actions are sending to middle layer and applying corresponding data to database. The data are transferred back to frontend and transformed into different marks on the map through Google map API.

5.4 International Relation Network Exhibition

The visualization interface can exhibit the international constructed from news. It also provides functions that allow users to check details of network. Through the interface, users can acquire a clear and comprehensive view of international network. Analysis can be

carried on the interface as well.

Fig 5-4 is showing important countries and hotspots. It based on BaiduNewsHit data. Size of red circle indicates BaiduNewsHit of the country. From the figure based on data with no time limitation, we can see that hotspots form a line across the world. It comes from East Asia, and then passes Central Asia, Middle East, reaches Western Europe, at last it ended in America. Countries far from the line gain less attention. Circles in Africa and Latin America are small and scattered.

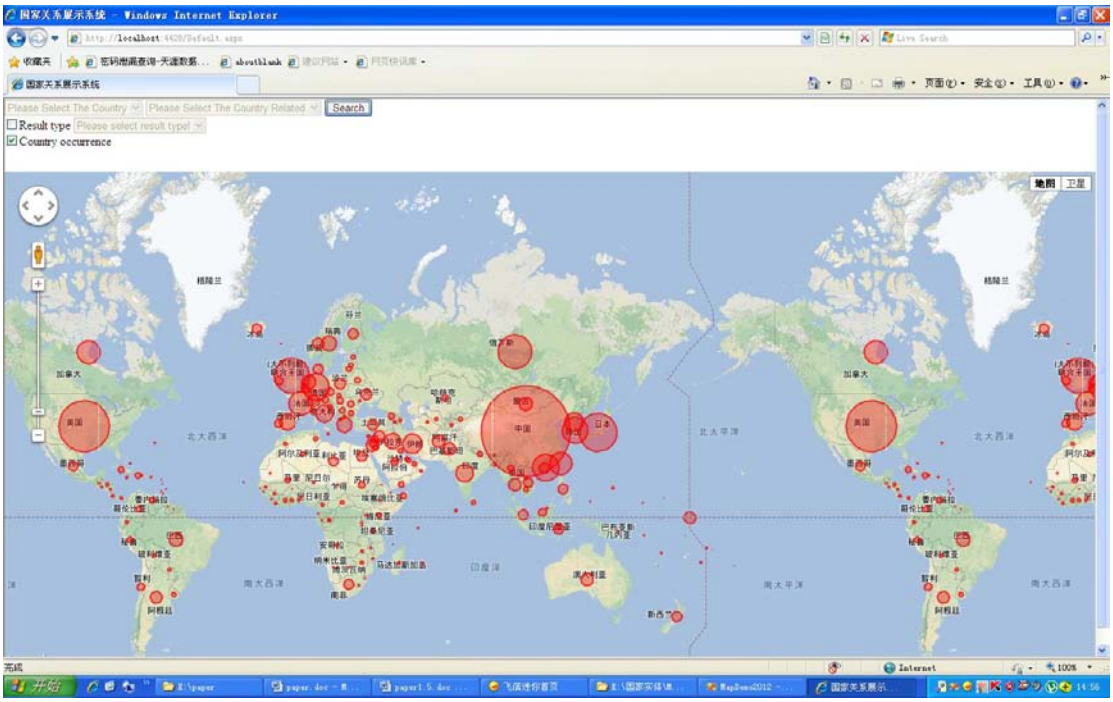


Fig 5-4: Hotspots of world

We also show top ten countries with high BaiduNewsHit in Table 5.1. All of them are close to the line.

Table 5.1 Top ten countries with high BaiduNewsHit

No time limitation		September 2012		October 2012	
country	Hits ×104	country	Hits ×104	country	Hits ×104
China	2110	China	15.4	China	61.7
America	631	Japan	10.4	America	9.9
Japan	374	America	9.45	Japan	6.88
Hong Kong	248	Tai Wang	2.73	Hong Kong	3.54
Britain	172	Hong Kong	2.67	Tai Wang	2.87
South Korea	169	Britain	2.44	South Korea	2.85
Tai Wang	168	South Korea	2.31	Britain	2.46

Germany	117	Germany	1.87	Germany	1.8
Russia	113	France	1.59	Spain	1.76
France	109	Iran	1.52	France	1.66

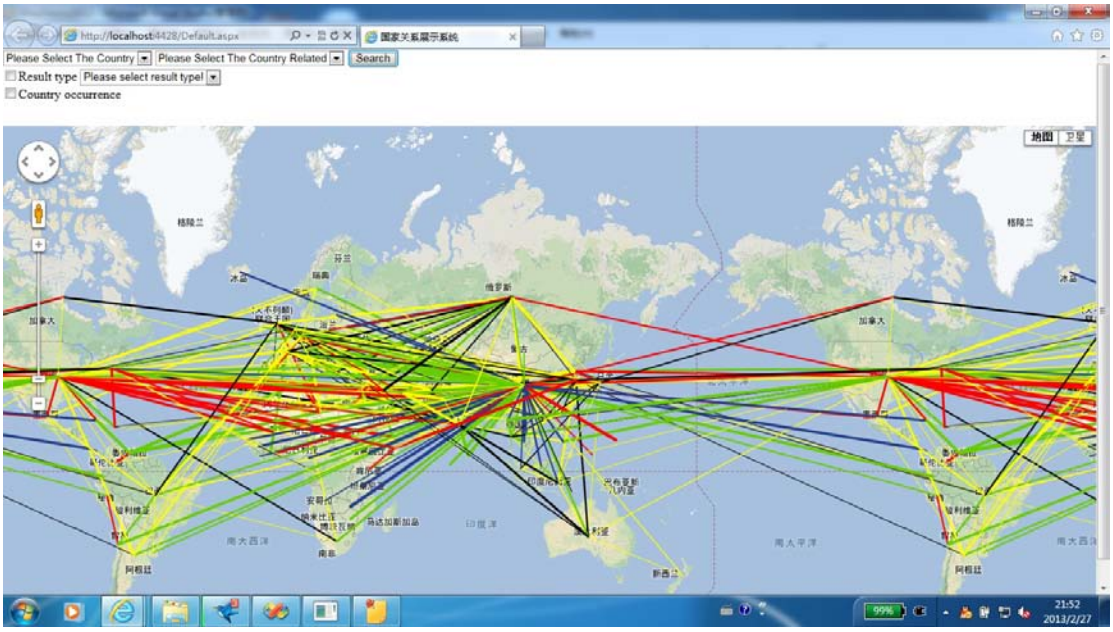


Fig 5-5: Overall international network

Fig 5-5 is the international network composed of all 400 country pairs. Different relations are represented by lines of different colors as in table 5.2. From the figure, we can see some countries with lots of lines, such as China, Russia, and America. These countries are critical countries in international relations. We also find an interesting phenomenon that America has most conflict relations. Is America really the biggest evil country?

Table 5.2 Relations and colors

relation	color
Confrontation	black
Neutral	yellow
Conflict	red
Subordination	green
Cooperation	blue

Through the international network of all countries, we can grasp the general view of international phenomenon. However, the overall network is a little complex for detailed research. In the interface, we also provide functions that allow users to see different aspects

of international network. With these functions, some interesting results are obtained as well. Following figures about interface are all based on data with no time limitation if there is no special explanation.

Fig 5-6 shows relation of America. We can see intuitively that America focus more on Central Asia, Middle East and East Asia. Enemies of it are mainly concentrated in two areas: Central Asia and Gulf of Aden.

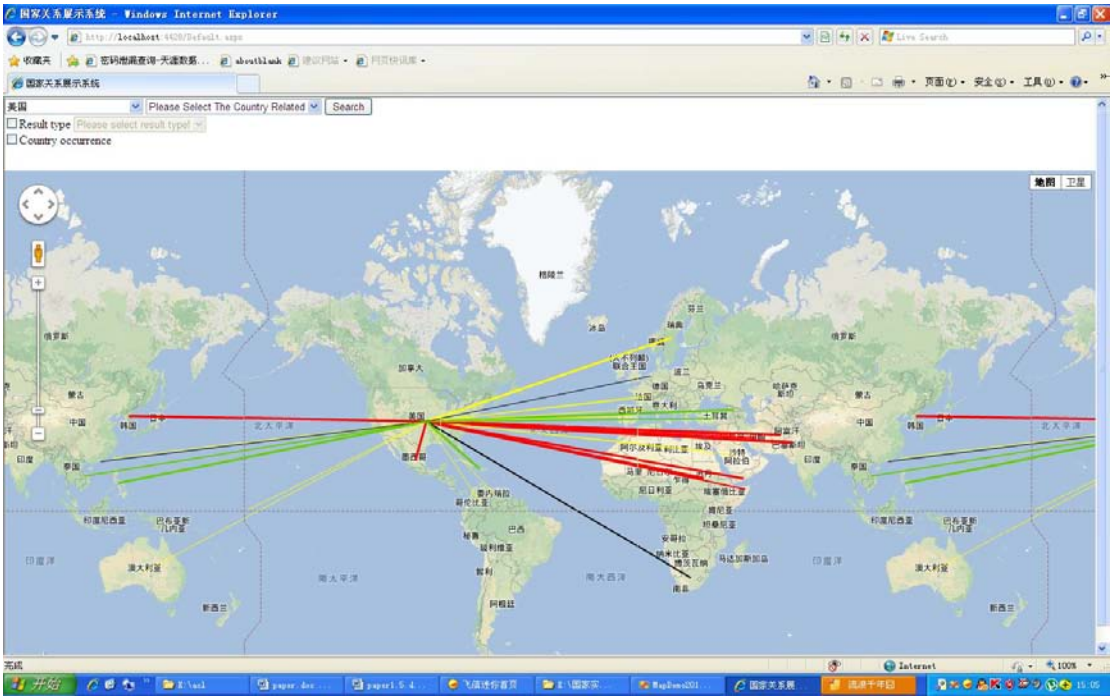


Fig 5-6: Relations of America

We also show the relation between America and Britain through the interface. Although their relation is neutral, they have many common concerns, especially two common enemies which are Libya and Iran. It indicates America and Britain may come closer for common benefit. In future work, we hope to design method that can predict future tendency of relations based on current relations.

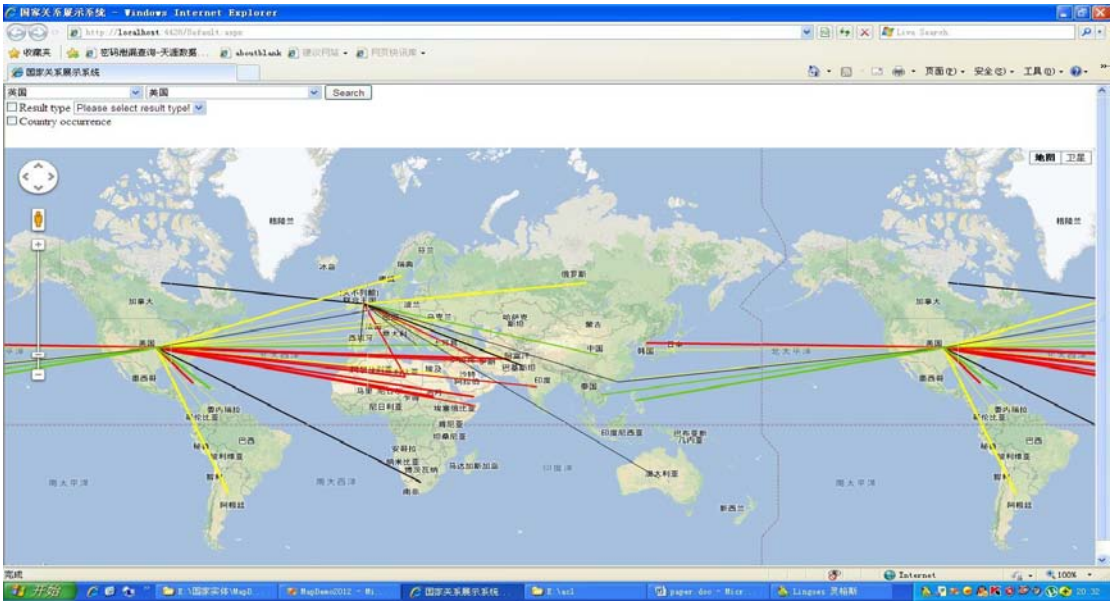


Fig 5-7: Relations between America and Britain

International network of different relations have interesting features as well. Take relation of cooperation and subordination as an example. Cooperation relations form a net while subordination relations form a tree. Nodes in cooperation network have relatively average node degree and the network has high connectivity. Subordination network has a few root nodes and lots of leaf nodes. Leaf nodes only have relations with their root. It can be explained by that cooperation relations are equal and friend of A can also be friend of B. However, in subordination network, one master can have several servants but one servant cannot serve too many masters. Cooperation and subordination relations of the other time periods have similar features. The interesting differences between the two relation networks also supports why we should separate relations into proper types.

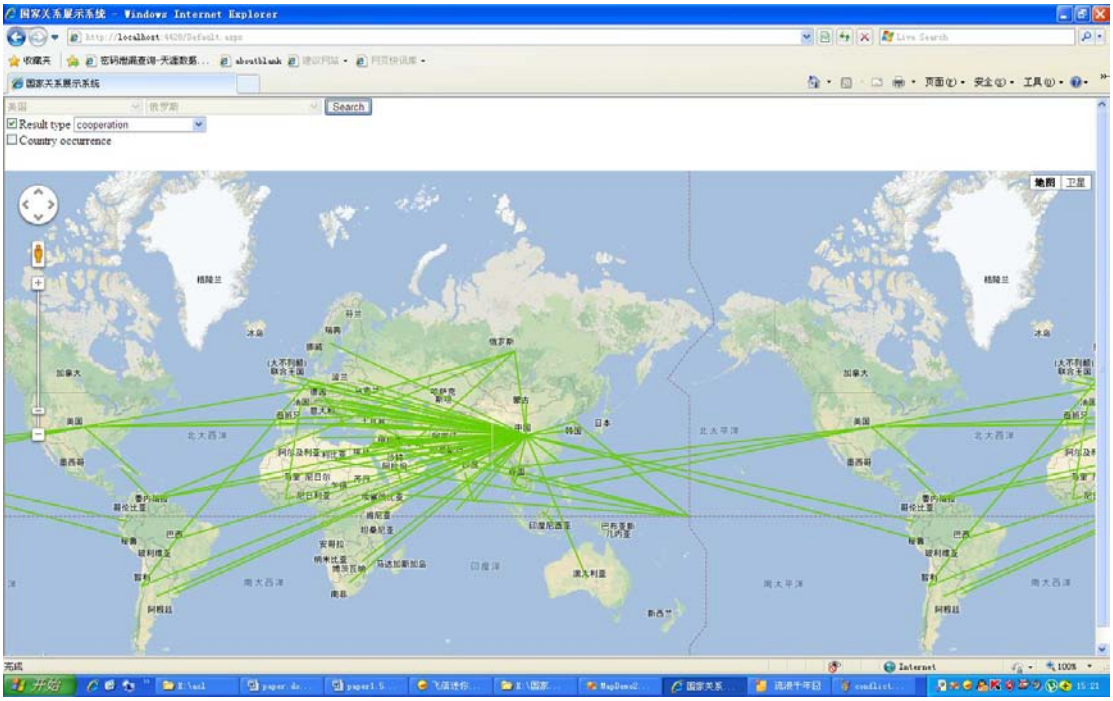


Fig 5-8: Relation of cooperation

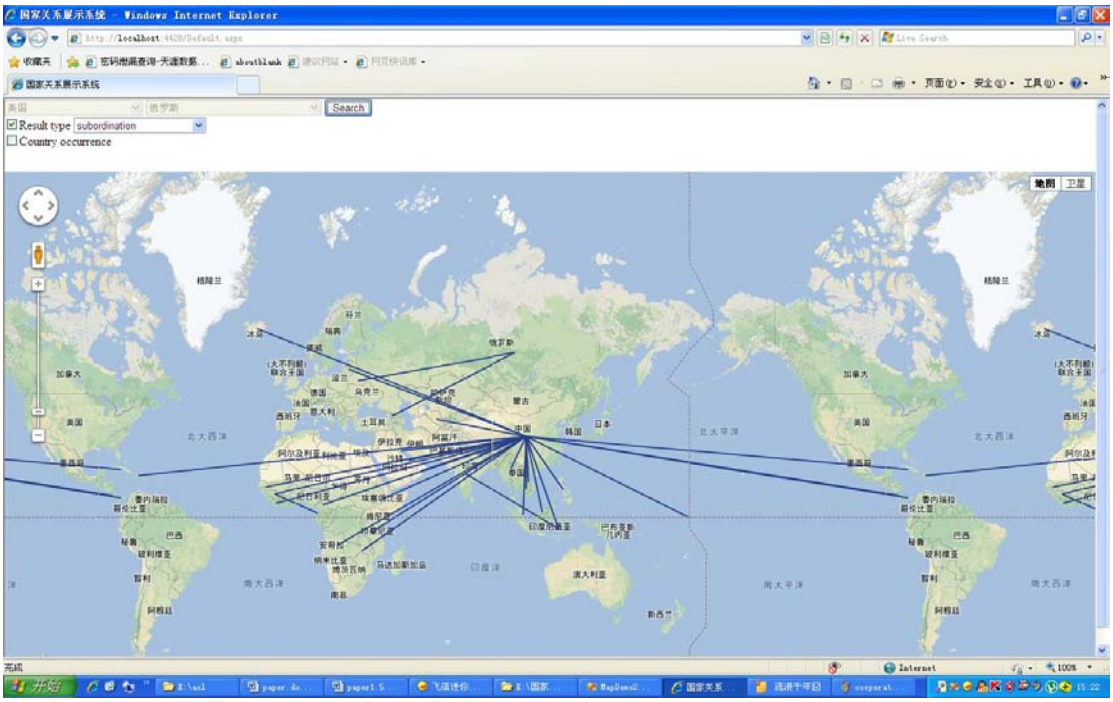


Fig 5-9: Relation of subordination

We also show figures of the other relations.

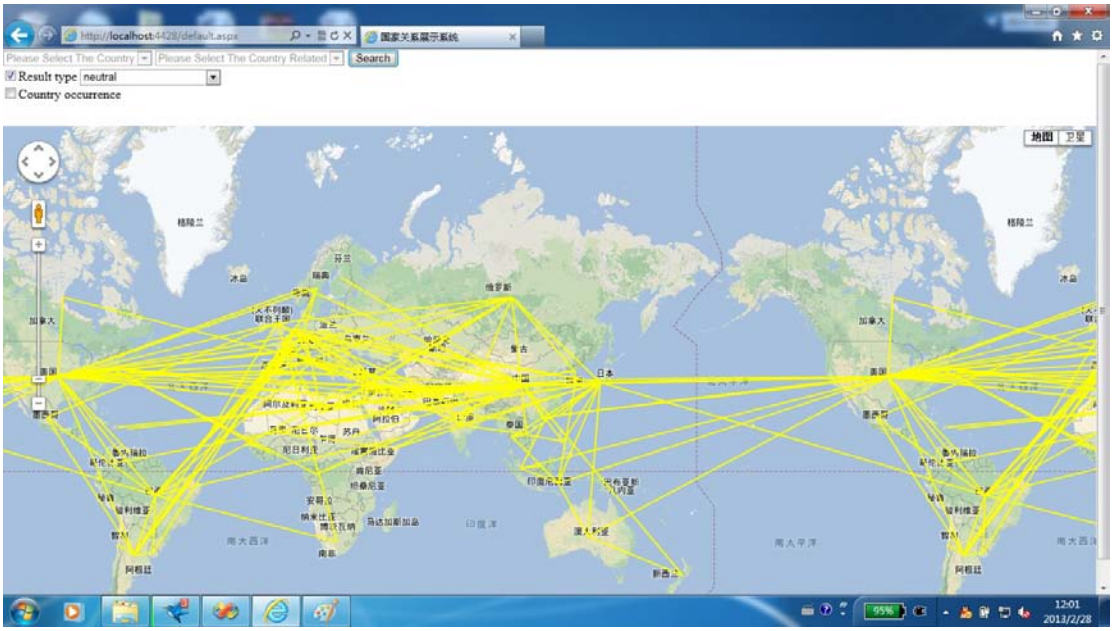


Fig 5-10: Relation of neutral

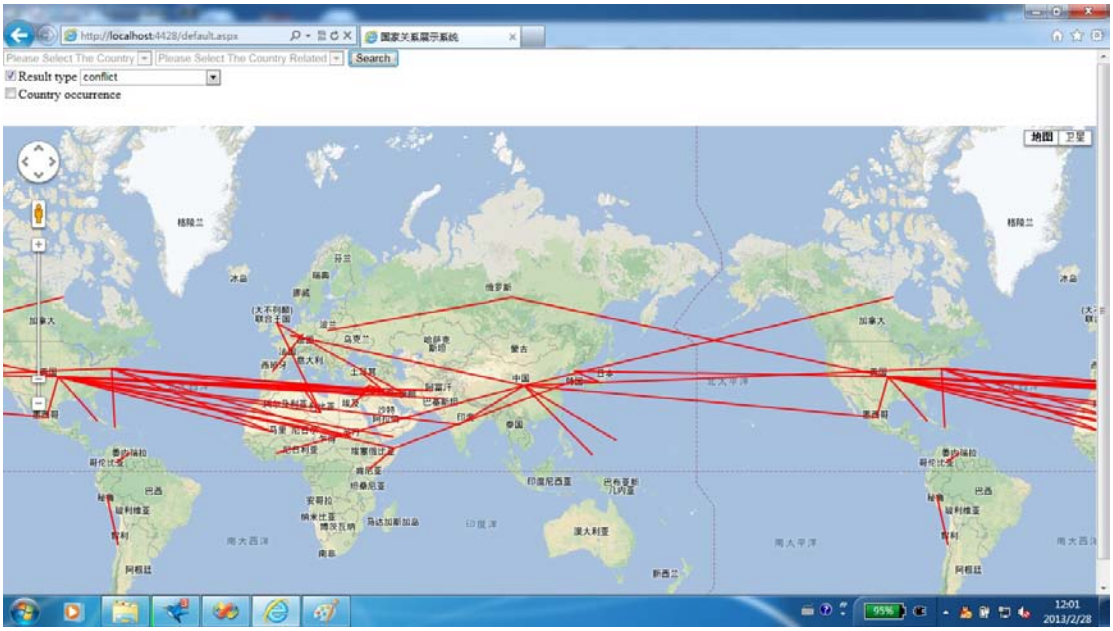


Fig 5-11: Relation of conflict

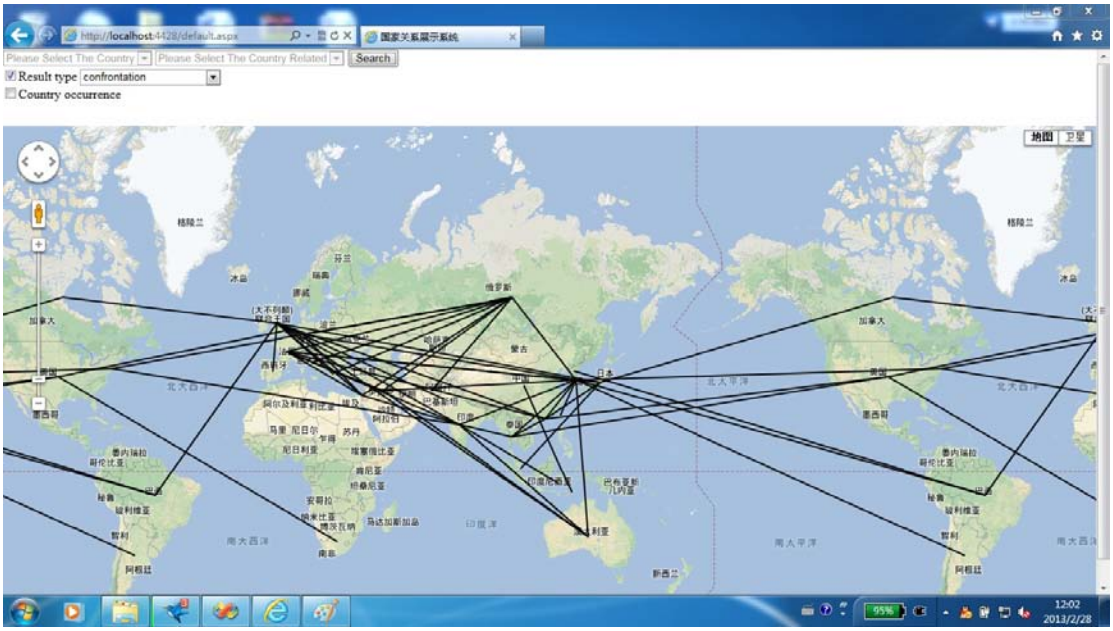


Fig 5-12: Relation of confrontation

From figures of different relations, we can find main distributions of each relation. For example, conflict relations are mainly concentrated with powerful countries, such as America, China and Russia. We also find that the five relations have different characteristics. Take the triadic closure as an example. Triadic closure is an interesting phenomenon of social network [90]. Some researchers consider it reflects transitivity of relation [91, 92]. If A and B have a common friends C, it is more likely that A and B will become friends through C which means they become a triadic closure.

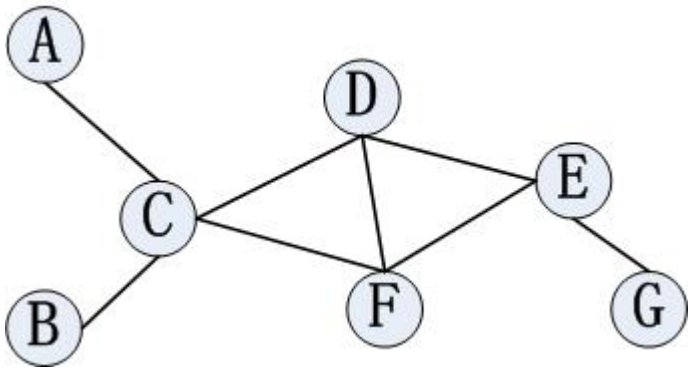


Fig 5-13: Triadic closure

In Fig5-13, nodes C-D-F and D-E-F form two triadic closures. They can be used to analyze the construction and evolution of network based on the view of structural balance

[93-95]. We count the number of triadic closure of each relation.

Table 5.3 Triadic closure of each relation

Relation	Number of pairs	Number of triadic closure
Confrontation	54	23
Neutral	172	112
Conflict	52	4
Subordination	33	0
Cooperation	89	12

From the table, we can see that neutral relation has most triadic closure while subordination and conflict has fewer. It may indicate effect of different relations in international communications and can be used to explain the evolution of international network.

Chapter 6 Conclusions and Future Work

In this thesis, we introduce the international network and show our method of constructing international network against disadvantages of current work. We define international network with its nodes and relations first, five new types of relations are proposed with their features. Based on the framework, we propose a prototype system first to test and verify our method.

In order to improve the defects of prototype system, we propose new sentiment analysis method based on CRF and collect news of three time periods through Chinese search engineer (baidu news) to expand data source. We analyze features of the data and explain the international network. Finally, our system achieves constructing international network from news automatically. In addition, an interface based on Google map is developed to show different aspect of the international network more vivid. Through the interface, we also have obtained some interesting results such as different characteristics of cooperation network and subordination network.

The main achievements are listed as follows:

(1) We achieve the goal of constructing international relation network from news. Based on related research about social network and interpersonal relationship, we point that intensity, quality and status are three important features of international relations. They affect the structure and evolution of networks in different way. We design several methods to extract these features from texts and construct the international relation network. Results of experiments show satisfied performance. Our method expands data source from structured data to unstructured texts and provide an efficient way to solve the problem of information

explosion.

(2) Recognizing sentiment of relations between entities. Sentiment between countries reflects quality of relations which is an important feature of international relations. We design a method based on CRF to recognize sentiment of relations in Chinese. Through entities recognition and extraction, sentiment related region detection and sentiment determination; we can obtain sentiment between entities on sentence level. We compare different algorithm based on different principle (rule and machine learning) and different related region as well. The algorithm using CRF (conditional random fields) model based on syntactic dependency tree acquires best result.

(3) Design and realization of visual interface. We analyze defects in current visualization approaches of international relation network and propose a new method combined with GIS. The interface uses Google Maps as substrate, integrates Google Maps API, Mysql database and IIS web service. It achieves international relation network visualization on electronic maps. The interface also provides some research functions that allow users to search interesting details and see different aspects of the international network more clearly. Through this interface, we analyze the constructed international relation network and obtain some interesting discoveries about hot spots and characters of network structure.

As future work, some more detailed research could be carried on. First, we hope more study about the structure feature and the evolution of international network, in order to explain why the structure of international network like this and how it will be. Second, we hope to search method of predicting development of relations between countries, which will be more useful for users.

References

[1] Winerman L. Social networking: Crisis communication. Nature, 2009, 457(7228): 376-378.

[2] Berry, Michael W., ed. Survey of Text Mining I: Clustering, Classification, and Retrieval. Vol. 1. Springer, 2004.

[3] Maoz Z, Terris L G, Kuperman R D, et al. International relations: A network approach. New directions for international relations: confronting the method-of-analysis problem, 2005: 35-64.

[4] Maoz Z. Networks of Nations: The evolution, structure, and effects of international networks, 1816-2001. UK: Cambridge University Press, 2010

[5] Renato Corbetta. Social Networks, Conflict and Cooperation In International Politics. The 2007Annual Meeting of the American Political Science Association . 2007

[6] Li KW, Hipel KW and Kilgour DM. Preference uncertainty in the graph model for conflict resolution. IEEE Transactions on System Man and Cybernetics Part A-System amd Humans, Vol.34, no.4, pp.507-520, 2004

[7] Liu B. Web data mining: exploring hyperlinks, contents, and usage data. Springer, 2007.

[8] Xu Haiyan, Hipel Keith W and Kilgour D Marc. Matrix Representation of Solution Concepts in Multiple-Decision-Maker Graph Models. IEEE Transactions on System Man and Cybernetics Part A-System amd Humans, Vol.39, no.1, pp.96-108, 2009

[9] Heaney M T, McClurg S D. Social Networks and American Politics Introduction to the Special Issue. American Politics Research, 2009, 37(5): 727-741.

[10] Emilie M, Hafner-Burton M K. Network analysis for international relations.

International Organization, 2009, 63: 559-92.

[11] Hafner-Burton E, Montgomery A. Globalization and the social power politics of international economic networks. Available at SSRN 1306648, 2008.

[12] Ohanyan A. Policy Wars for Peace: Network Model of NGO Behavior. International Studies Review, 2009, 11(3): 475-501.

[13] Dorussen H, Ward H. Trade networks and the politics of cooperation and conflict. Unpublished manuscript, 2008.

[14] Davey A. Statistical power analysis with missing data: A structural equation modeling approach. Routledge, 2009.

[15] Little R J A, Rubin D B. Statistical analysis with missing data. New York: Wiley, 1987.

[16] http://www.correlatesofwar.org/

[17] Scott and John, Social Network Analysis: A Handbook. 2nd ed. London, UK: SAGE Publications, 2000.

[18] Carrington P J, Scott J, Wasserman S (Eds.). Models and Methods in Social Network Analysis. New York: Cambridge University Press, 2005.

[19] Vega-Redondo F. Complex Social Networks. New York: Cambridge University Press, 2007.

[20] Mika P. Flink: semantic web technology for the extraction and analysis of social networks. Web Semantics: Science, Services and Agents on the World Wide Web, 2005, 3(2):211–223

[21] Hämmerli, R.Gattiker amd R.Weyermann. Conflict and cooperation in an actors' network of Chechnya based on event data. Journal of Conflict Resolution 2006, vol 50, pp.159–175, 2006

[22] Kautz H, Selman B, Shah M. Referral Web: combining social networks and collaborative filtering. Communications of the ACM, 1997, 40(3): 63-65.

[23] Yutaka Matsuo, Junichiro Mori and Masahiro Hamasaki. POLYPHONET: an advanced social network extraction system from the web. Proceedings of the 15th international conference on World Wide Web, pp. 262-278, Edinburgh, Scotland, 2006.

[24] Weijie YANG, Ruwei DAI and Xia CUI. A New Kind of Social Network and Its Application in Text Mining. Proceedings of the Fourth National Conference of Information Retrieval and Content Security Vol.2, 2008

[25] Marlow C, Byron L, Lento T, et al. Maintained relationships on Facebook. Retrieved February, 2009, 15: 2010.

[26] Kazienko P, Musial K, Kajdanowicz T. Multidimensional Social Network in the Social Recommender System. IEEE Transactions on System Man and Cybernetics Part A-System and Humans, Vol.41, no.4, pp.746-759, 2011.

[27] J. Kunegis, A. Lommatzsch, and C. Bauckhage. The Slashdot Zoo: Mining a social network with negative edges. In Proc. 18th WWW, pp. 741–750, Madrid, Spain 2009.

[28] http://ediplomacy.afp.com/

[29] Granovetter M. Getting a job: A study of contacts and careers. University of Chicago Press, 1995.

[30] Granovetter M S. The strength of weak ties. American journal of sociology, 1973: 1360-1380.

[31] M. J. Brzozowski, T. Hogg, and G. Szabó. Friends and foes: ideological social networking. In Proc. 26th CHI, pp. 817-820, Florence, Italy, 2008.

[32] Yanyan Shao. Testing and Adjustment of Interpersonal Relationship. Culture Press of Shang Hai,1988

[33] C. Faloutsos, K. S. McCurley and A. Tomkins. Fast discovery of connection subgraphs. In Proc. ACM SIGKDD 2004, pp. 118-127, Seattle, USA, 2004.

[34] P. Knees, E. Pampalk and G. Widmer. Artist classification with web-based data. In 5th

International Conf. on Music Information Retrieval (ISMIR), Barcelona, Spain, 2004.

[35] C. D. Manning and H. Sch¨utze. Foundations of Statistical Natural Language Processing. The MIT Press, London, 2002.

[36] C., Zong. Statistical Natural Language Processing. Tsinghua University Press, 2008.

[37] R. Sun. Basic Ecology. Higher Education Press,2007.

[38] Schrodt, Philip A and Shannon G Davis et al. Political science: KEDS – a program for the machine coding of event data. Social Science Computer Review vol.12, no.3, pp.61–588, 1994

[39] IDEA. http://vranet.com/IDEA/

[40] Bond Doug, Joe Bond and Churl Oh et al. Integrated data for events analysis (IDEA): An event typology for automated event data development. Journal of Peace Research vol.40, no.6, pp.733-45, 2003.

[41] C. Quan and F. Ren. A blog emotion corpus for emotional expression analysis in Chinese. Computer Speech and Language, v.24 n.4, pp.726-749, 2010.

[42] F. Ren and C. Quan. Linguistic-Based Emotion Analysis and Recognition for Measuring Consumer Satisfaction - An Application of Affective computing. Information Technology and Management, 2012, DOI: 10.1007/s10799-012-0138-5

[43] H. Xu, L.Sun, T.Yao. The third Chinese opinion analysis evaluation (COAE2011). Institute of Computing Technology, China Academy of Science,2011

[44] T. Wilson, J. Wiebe , and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, p.347-354, Vancouver, British Columbia, Canada, 2005.

[45] R. McDonald, K. Hannan and T. Neylon et al. Structured models for fine-to-coarse sentiment analysis. In Proceedings of the Association for Computational Linguistics

(ACL), pp. 432-439, Prague, Czech Republic, 2007

[46] Kim, Soo-Min and Eduard Hovy. Identifying and Analyzing Judgment Opinions. Proceedings of HLT/NAACL-2006, pp. 200-207, New York City, NY, 2006

[47] Kim, Soo-Min and Eduard Hovy. Crystal: Analyzing predictive opinions on the web. In Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL) , pp, 1056-1064, Prague, Czech Republic 2007.

[48] Tetsuji Nakagawa, Kentaro Inui and Sadao Kurohashi. Dependency tree-based sentiment classification using CRFs with hidden variables. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp.786-794, Los Angeles, California, 2010.

[49] Binyang Li, Lanjun Zhou, Shi Feng, and Kam-Fai Wong. A Unified Graph Model for Sentence-based Opinion Retrieval. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 1367–1375,Uppsala, Sweden, 2010.

[50] Fuji Ren. Affective Information Processing and Recognizing Human Emotion. Electronic Notes in Theoretical Computer Science, Vol.225, No.2009, pp.39-50, 2009.

[51] Fuji Ren and B. David. Advanced Information Retrieval. Electronic Notes in Theoretical Computer Science, Vol.225, No.1, pp.303-317, 2009.

[52] Faust, Katherine, Karin E.Willert and D. Rowlee et al. Scaling and statistical models for affiliation networks: Patterns of participation among Soviet politicians during the Brezhnew era. Social Networks vol.24, pp.231-59, 2002.

[53] Moore, Spencer, Eugenia Eng, and Mark Daniel. International NGOs and the role of network centrality in humanitarian aid operations: A case study of coordination during the 2000 Mozambique floods. Disasters vol.27, no.4, pp.305-18, 2003

[54] Current affairs data. http://www.banyuetan.org/

[55] M. Harada, Sh. Sato and K. Kazama. Finding authoritative people from the web. In Proc. Joint Conference on Digital Libraries (JCDL), pp. 306-313, Tucson, Arizona, USA, 2004.

[56] Wanxiang Che, Zhenghua Li and Ting Liu. LTP: A Chinese Language Technology Platform. In Proceedings of the Coling 2010: Demonstrations. pp13-16, Beijing, China, 2010.

[57] K.Soo-Min and E.Hovy. Determining the Sentiment of Opinions. Proceedings of COLING-04. pp. 1367-1373. Geneva, Switzerland(2004)

[58] X. Kang and F.Ren. Predicting Complex Word Emotions and Topics through a Hierarchical Bayesian Network. China Communications, Vol.9No.3, pp.99-109 (2012)

[59] Jiang Pei-Lin, Wang Fei and Lin Fu-Ji. Semi-Automatic Complex Emotion Categorization and Ontology Construction from Chinese Knowledge. China Communications, Vol.9No.3, pp.28-37 (2012)

[60] LiuGongshen, HeWenlei, ZhuJie, et al. Feature Repre-sentation Based on Sentimental Orientation Classification. China Communications, Vol.8No.3, pp.90-98 (2011)

[61] Tetsuji Nakagawa , Kentaro Inui and Sadao Kurohashi. Dependency tree-based sentiment classification using CRFs with hidden variables. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp.786-794, Los Angeles, California(2010)

[62] John D. Lafferty , Andrew McCallum and Fernando C. N. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Proceedings of the Eighteenth International Conference on Machine Learning, pp.282-289 (2001)

[63] CRF++: Yet Another CRF toolkit. http//sourceforge.net

[64] H.Duan, Y.Zheng. A Study on Features of the CRFs-based Chinese Named Entity Recognition. International Journal of Advanced Intelligence,Vol.3No.2,

pp.287-294(2011)

[65] Akamine S, Kawahara D, Kato Y, et al. WISDOM: a web information credibility analysis system. Proceedings of the ACL-IJCNLP 2009 Software Demonstrations. Association for Computational Linguistics, 2009: 1-4.

[66] Lee R, Kitayama D, Sumiya K. Web-based evidence excavation to explore the authenticity of local events. Proceedings of the 2nd ACM workshop on Information credibility on the web. ACM, 2008: 63-66.

[67] http://news.baidu.com/

[68] Rübel O, Ahern S, Bethel E, et al. Coupling visualization and data analysis for knowledge discovery from multi-dimensional scientific data. Procedia computer science, 2010, 1(1): 1757-1764.

[69] Upson C, Faulhaber Jr T A, Kamins D, et al. The application visualization system: A computational environment for scientific visualization. Computer Graphics and Applications, IEEE, 1989, 9(4): 30-42.

[70] Readings in information visualization: using vision to think. Morgan Kaufmann, 1999.

[71] van Ham F, van Wijk J J. Interactive visualization of small world graphs. Information Visualization, 2004. INFOVIS 2004. IEEE Symposium on. IEEE, 2004: 199-206.

[72] Kamada T, Kawai S. An algorithm for drawing general undirected graphs. Information processing letters, 1989, 31(1): 7-15.

[73] Sugiyama K, Tagawa S, Toda M. Methods for visual understanding of hierarchical system structures. Systems, Man and Cybernetics, IEEE Transactions on, 1981, 11(2): 109-125.

[74] Fruchterman T M J, Reingold E M. Graph drawing by force‐directed placement. Software: Practice and experience, 1991, 21(11): 1129-1164.

[75] Battista G D, Eades P, Tamassia R, et al. Algorithms for drawing graphs: an annotated

bibliography. Computational Geometry, 1994, 4(5): 235-282.

[76] Kraak M J, MacEachren A. Visualization for exploration of spatial data. International Journal of Geographical Information Science, 1999, 13(4): 285-287.

[77] Blok, Cornelia Adriana. Dynamic visualization variables in animation to support monitoring of spatial phenomena. Utrecht University, 2005.

[78] van de Vlag D E. Modeling and visualizing dynamic landscape objects and their qualities. ITC, 2006.

[79] Ostermann, Frank. Modeling, analyzing, and visualizing human space appropriation. Diss. Universität Zürich, 2009.

[80] Lee B, Plaisant C, Parr C S, et al. Task taxonomy for graph visualization. Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization. ACM, 2006: 1-5.

[81] Kraak, Menno-Jan. The role of the map in a Web-GIS environment. Journal of Geographical Systems 6.2 (2004): 83-93.

[82] Boulos, MN Kamel, et al. Web GIS in practice IX: a demonstration of geospatial visual analytics using Microsoft Live Labs Pivot technology and WHO mortality data. International journal of health geographics 10.1 (2011): 19.

[83] Lu, Xiaolin. An investigation on service-oriented architecture for constructing distributed web gis application. Services Computing, 2005 IEEE International Conference on. Vol. 1. IEEE, 2005.

[84] Liu, Rui, and Yongqi Ge. WebGIS architecture research and design for forestry resources application. Computer Science & Education (ICCSE), 2013 8th International Conference on. IEEE, 2013.

[85] Yuan, Wei, Mei Hong, and Dong-mei WEI. WebGIS Design and Implementation Based on B/S Mode. Computer Technology and Development 8 (2008): 004.

[86] WANG, Ping, and Hua-ping JIA. Study and development of application system based on combination of C/S and B/S. Computer and Information Technology 1 (2006): 015.

[87] Tan, Xiaojun, et al. Integration WebGIS with AJAX and XML based on google maps. Intelligent Networks and Intelligent Systems, 2008. ICINIS'08. First International Conference on. IEEE, 2008.

[88] Google maps api. www.google.com/apis/maps /

[89] Rapoport A. Spread of information through a population with socio-structural bias: I. Assumption of transitivity. The bulletin of mathematical biophysics, 1953, 15(4): 523-533.

[90] Huang, Hong, et al. Mining triadic closure patterns in social networks. Proceedings of the companion publication of the 23rd international conference on World wide web companion. International World Wide Web Conferences Steering Committee, 2014.

[91] Zhao, Jiayun, and S. Ram. Examining the evolution of networks based on lists in twitter. Internet Multimedia Systems Architecture and Application (IMSAA), 2011 IEEE 5th International Conference on. IEEE, 2011.

[92] Opsahl, Tore. Triadic closure in two-mode networks: Redefining the global and local clustering coefficients. Social Networks 35.2 (2013): 159-167.

[93] Kossinets, Gueorgi, and Duncan J. Watts. Empirical analysis of an evolving social network. Science 311.5757 (2006): 88-90.

[94] Lou, Tiancheng, et al. Learning to predict reciprocity and triadic closure in social networks. ACM Transactions on Knowledge Discovery from Data (TKDD) 7.2 (2013): 5.

[95] Klimek, Peter, and Stefan Thurner. Triadic closure dynamics drives scaling laws in social multiplex networks. New Journal of Physics 15.6 (2013): 063008.