UNIVERSITY OF TURKU

ABSTRACT

	Bachelor's thesis
×	Master's thesis
	Licentiate's thesis
	Doctor's thesis

Subject	Information Systems Science	Date	27th June, 2019
Author(s)	Yajing Wang	Student number	609798
		Number of pages	78
Title	A comparative study of Chinese and European Internet companies' privacy policy based on knowledge graph		
Supervisor(s)	Prof. Reima Suomi		

Privacy policy is not only a means of industry self-discipline, but also a way for users to protect their online privacy. The European Union (EU) promulgated the General Data Protection Regulation (GDPR) on May 25th, 2018, while China has no explicit personal data protection law. Based on knowledge graph, this thesis makes a comparative analysis of the Chinese and European Internet companies' privacy policies, and combines with the relevant provisions of GDPR, puts forward suggestions on the privacy policy of Internet companies, so as to solve the problem of personal information protection to a certain extent.

Firstly, this thesis chooses the process and methods of knowledge graph construction and analysis. The process of constructing and analyzing the knowledge graph is: data preprocessing, entity extraction, storage in graph database and query. Data preprocessing includes word segmentation and part-of-speech tagging, as well as text format adjustment. Entity extraction is the core of knowledge graph construction in this thesis. Based on the principle of Conditional Random Fields (CRF), CFR++ toolkit is used for the entity extraction. Subsequently, the extracted entities are transformed into ".csv" format and stored in the graph database Neo4j, so the knowledge graph is generated. Cypher query statements can be used to query information in the graph database.

The next part is about comparison and analysis of the Internet companies' privacy policies in China and Europe. After sampling, the overall characteristics of the privacy policies of Chinese and European Internet companies are compared. According to the process of constructing knowledge graphs mentioned above, the "collected information" and "contact us" parts of the privacy policy are used to construct the knowledge graphs.

Finally, combined with the relevant content of GDPR, the results of the comparative analysis are further discussed, and suggestions are proposed. Although Chinese Internet companies' privacy policies have some merits, they are far inferior to those of European Internet companies. China also needs to enact a personal data protection law according to its national conditions.

This thesis applies knowledge graph to the privacy policy research, and analyses Internet companies' privacy policies from a comparative perspective. It also discusses the comparative results with GDPR and puts forward suggestions, and provides reference for the formulation of China's personal information protection law.

Key words	Privacy policy, Knowledge graph, Entity extraction, Conditional Random Fields
Further information	

UNIVERSITY OF TURKU
Turku School of Economics

A COMPARATIVE STUDY OF CHINESE AND EUROPEAN INTERNET COMPANIES' PRIVACY POLICY BASED ON KNOWLEDGE GRAPH

Masters Thesis
in Information Systems Science

Author:
Yajing Wang

Supervisor:
Prof. Reima Suomi

27.06.2019
Turku

The originality of this thesis has been checked in accordance with the University of Turku quality assurance system using the Turnitin OriginalityCheck service.

Table of contents

List of figures

List of tables

List of abbreviations

EU	European Union
GDPR	General Data Protection Regulation
CRF	Conditional Random Fields
HMM	Hidden Markov Models
MEMM	Maximum Entropy Markov Models
POS tagging	Part-of-speech tagging

1 INTRODUCTION

1.1 Background

In the information era, data is money, providing opportunities for business (Ohlhorst, 2012). However, personal information is included in the data collected by companies and is in the risk of being offended. AT&T Company buys "Internet Preferences" which includes personal browsing data from subscribers for $20 a month. Datacoup Company buys monthly activity data and credit card usage from users on Facebook. (Savage & Waldman, 2015) Users are worried that the companies abuse their personal information and are looking for good protection (Attaran & VanLaar, 2002). How to ensure the privacy and security of users' information is becoming a hot issue.

Protecting personal information is a trend all over the world. The EU's (European Union) GDPR (General Data Protection Regulation) was formally enforced on May 25th, 2018. It is reported that on May 28th, some American enterprises such as Facebook and Google became the first defendants under the GDPR (Hill, 2018). On September 10th, 2018, China proposed to establish the Personal Information Protection Law in the 13th Legislative planning of the Standing Committee of the National People's Congress (NPC China, 2018).

As early as 1999, Gindin (1999) put forward creating an online privacy policy to protect online users' privacy. Privacy policy is one of the basic elements of loyalty to a website (Flavián & Guinalíu , 2006). Privacy policy is not only a way of corporate self-discipline, but also a means to protect users' private information (Zhang, 2017). It is necessary to study the rationality of the privacy policy clauses. Good privacy clauses not only protect users' privacy and security, but also urge enterprises to assume the responsibility of protecting users' privacy.

1.2 Main concepts

1.2.1 Internet company

The Internet company here refers to "the dot-com company, is a company that does most of its business on the Internet" (TheFreeDictionary). Internet companies can be categories as search engine like Google, comprehensive information portal like Yahoo, instant messaging like WhatsApp, and e-commerce like Amazon.

One feature of Internet companies is they collect a large amount of data including users' personal information every day to support their business, so their privacy policies are worth studying.

1.2.2 Privacy policy

The privacy policy refers to the disclaimers for informing users that their personal data may be collected, used and shared with third parties and how the websites use the data, to inform users the information security issues when they use the website, so as to reach a consensus with users on privacy protection (Pollach, 2006). Privacy policy also stipulates that the company should assume the obligation to protect the lawful right and interests of users in personal information.

Normally, the website would set the privacy policy as a hyperlink and put it at the end of the webpage.

1.2.3 Knowledge graph

There is no formal and specific definition of knowledge graph. Paulheim (2017) concluded part of features of knowledge graph. According to him, knowledge graph is a semantic network composed of concepts, entities, events and their relationships (Li & Hou, 2017):

- Concepts refer to the conceptual representation of objective things formed in the process of people's understanding of the world, such as human, animal, organization, etc.
- Entities are specific things in the objective world, such as basketball player Kobe Bryant, Internet company Tencent and so on.
- Events are activities of the objective world, such as earthquake, trading and so on.
- Relationships describe the objective relationships among concepts, entities and events, such as the couple relationship between David Beckham and Victoria Beckham, the relationship between the concepts and sub-concepts like players and basketball players, etc.

Figure 1-1 (Ontotext) shows a simple example of knowledge graph. The circular, or shall be called node, represents different entities such as France. The line between the nodes stands for the relationships between the entities. The complicated data is stored as entities and their relationships in the semantic network. When we want to know some

information, it can be quickly called out. We can also easily get the information of other entities from the known entities by query language.

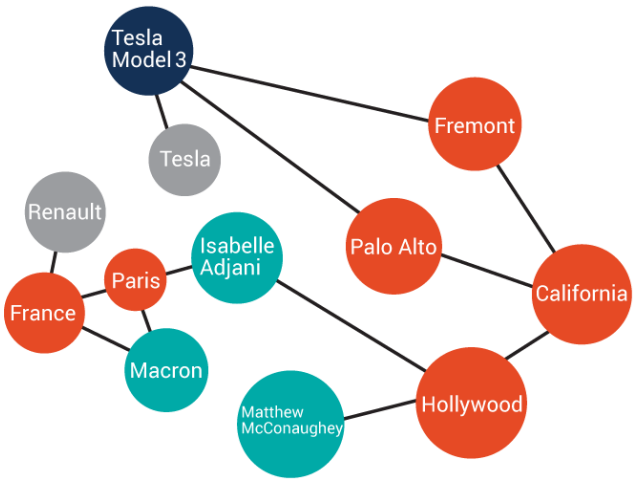


Figure 1-1 An example of knowledge graph (adopted from Ontotext)

The construction of knowledge graph is processing scattered structured, semi-structured and unstructured data from web and making them into structured data by technologies such as named entity recognition and knowledge fusion (Qi, Gao & Wu, 2017), which is convenient for the upper application system to analyze complex logical reasoning problems from the angle of the whole knowledge system. In this sense, knowledge graph benefits from the development of the Web, supported by Knowledge Representation (KR), Natural Language Processing (NLP), Web, Artificial Intelligence (AI) and many other aspects.

1.2.4 GDPR (General Data Protection Regulation)

In November 2012, the European Commission drafted an inclusive and cooperative General Data Protection Regulation (GDPR). In April 14th, 2016, the European Union held a meeting and adopted GDPR. On May 25th, 2018, the GDPR takes effect in the 28 member states of the European Union. There are 173 prefaces to the GDPR. The text is divided into 11 chapters and 99 articles (EU GDPR.ORG).

The GDPR is a regulation in EU law to protect the personal information and privacy of all individuals within the European Union (EU) and the European Economic Area (EEA). Its predecessor Data Protection Directive was created in 1995 (Voigt & Von dem Bussche, 2017). By unifying the regulation within the EU, GDPR enables citizens to control their own data and simplifies the regulatory environment for international business. In the official website of GDPR (EU GDPR.ORG), it shows “The regulation

will fundamentally reshape the way in which data is handled across every sector, from healthcare to banking and beyond". There are several clarifications and changes in the factors affecting the validity of consent in the GDPR (Lappalainen, 2017). Organizations need to ensure themselves are not caught out and face sanctions and high fines for non-compliances with GDPR (Tankard, 2016).

1.3 Motivation

This thesis aims to find the similarities and differences between the Chinese and the European Internet companies' privacy policy, and give suggestions with GDPR, to improve the privacy policy, making it both contribute to the business benefits and conform to the regulations.

- Why study privacy policy?
  A large amount of personal data is collected every day. Regulations require data controllers to inform their users about their data collection and processing procedures. One way to inform users is through the privacy policy. (Tesfay, Hofmann, Nakamura, Kiyomoto & Serna, 2018) On this basis, privacy policy is a good perspective to help to improve the security of personal data. Privacy policy restricts what information the enterprises can collect and how they would use it. If the privacy policy is fair enough, the users' privacy rights can be protected to some extent.
- Why choose Internet companies?
  Internet companies do most of its business on the Internet. Many of them make benefits by the visitors' flow. How to attract and retain customers relates tightly with their development. Data, especially personal data is significant for them. Additionally, the Internet has aroused many critical discussions because of its interactive nature. Its privacy, unsolicited e-mail, transaction security and pornography have been the hot topics in academic circles. (Cook & Coupey, 1998; Koprowski, 1995) In this sense, Internet companies is definitely a good example to study privacy policy.
- Why use Knowledge Graph?
  Privacy policy is usually too long for people to read in a short time (Liu, Wilson, Story, Zimmeck & Sadeh, 2018). It is also difficult to analyze a big sample of privacy policy by reading and comprehension. Privacy policies of a region, such as China, have strong consistency in structure and content, because they are formulated according to the relevant provisions of the laws of the region. However, these privacy policies are scattered in the websites of various enterprises, and unstructured text information makes it impossible for people to

get information from the whole angle. At this time, the knowledge graph can store the chaotic privacy policy text in the form of entity and relationship, and further analysis can be carried out by using query language to get specific information.

- Comparative study?
  The comparative method is a fundamental analytical method to describe the suggestive resemblances and dissimilarities among cases. The results by using a few cases to do the comparative study are affected heavily by the political environment (Collier, 1993). GDPR, which is famous of strictness, has come into force in European place, so is there differences of privacy policy between Europe and China? A comparative study can be taken to solve this question.

1.4 Research gap

The research gap of privacy policy is mainly on the research method. There have been many researches upon privacy policy. Pollach (2005; 2006; 2007) is an expert to study privacy policy. Nevertheless, his studies are mainly from an ethic perspective, and use statistical methods. Some Chinese writers such as Tang and Lai (2018) study the privacy policy from the comparative perspective, but they just stop at a level of text analysis. They usually choose two typical companies' privacy policies and have simple comparisons, which is not universal to find the gaps. The nature of privacy policy is text, so there are some scholars study it by Natural Language Processing (NLP) technology. For example, Story et al. (2019) frame the classification problem in privacy policies by NLP. However, their focus is the improvement of the NLP effects rather than finding the problems of specific privacy policies.

On the other hand, this study expands the domain knowledge graph. Although open knowledge graphs like YAGO (Suchanek, Kasneci & Weikum, 2007) and DBpedia (Lehmann et al., 2015) are very mature, the domain knowledge graph is in a need of development. The domain knowledge graph is the research focus in recent years. For example, the knowledge graph is constructed based on geoscience literature (Wang, Ma, Chen & Chen, 2018). Tennakoon, Zaki, Arnaout, Elbassuoni, El-Hajj and Al Jaberi (2019) have a research on biological knowledge graph construction, search, and navigation. Health knowledge graph is constructed from electronic medical records (Rotmensch, Halpern, Tlimat, Horng & Sontag, 2017). This study tends to construct privacy policy knowledge graphs from privacy policies of Chinese and European Internet companies. Thus, the domain knowledge graph is expanded in the field of privacy policy to some extent.

This study applies knowledge graph to the privacy policy research, and analyses Internet companies' privacy policies from a comparative perspective with GDPR. It takes account of both technology and practical relevance, which is the best of both worlds. This study tends to fill the research gap on three aspects:

- Applying knowledge graph to the privacy policy research.
  As far as I know, no literature constructs knowledge graphs of privacy policy. This study uses knowledge graph technologies to visualize the privacy policies, and analyzes the privacy policies by querying in the knowledge graphs.
- Expand the domain knowledge graph in the field of privacy policy.
  This study tends to construct privacy policy knowledge graphs from privacy policies of Chinese and European Internet companies. Thus, the domain knowledge graph is expanded in the field of privacy policy to some extent.
- Discussing the privacy policy with GDPR.
  GDPR is new, enforced on May 25th, 2018. So discussing the privacy policy with GDPR is in a trend of cutting edge. This thesis analyses Internet companies' privacy policies with GDPR, expecting to provide reference for the formulation of China's personal information protection law.

1.5 Research question

To provide suggestions for the Internet companies about their privacy policies, two main research questions and several sub questions are put forward as follows:

(1) How to construct privacy policy knowledge graphs of Chinese and European Internet companies?
   1) What are privacy policy and knowledge graph?
   2) How to choose good samples as data sources for the construction of privacy policy knowledge graph?
   3) What kind of knowledge graphs shall be constructed?
   4) Which methods can be used in the process of constructing knowledge graph, in order to get good results?

(2) How to compare and analyze privacy policies of Chinese and European Internet companies?
   1) What are the similarities and differences between privacy policies of Chinese and European Internet companies?
   2) What is the relevant content of GDPR for our comparative study?
   3) What suggestions can be provided according to the analytic results?

1.6 Main works and the structure

The main works of this study are constructing knowledge graphs of the privacy policy and discussing the comparative results with GDPR. Figure 1-2 shows the structure of this study.

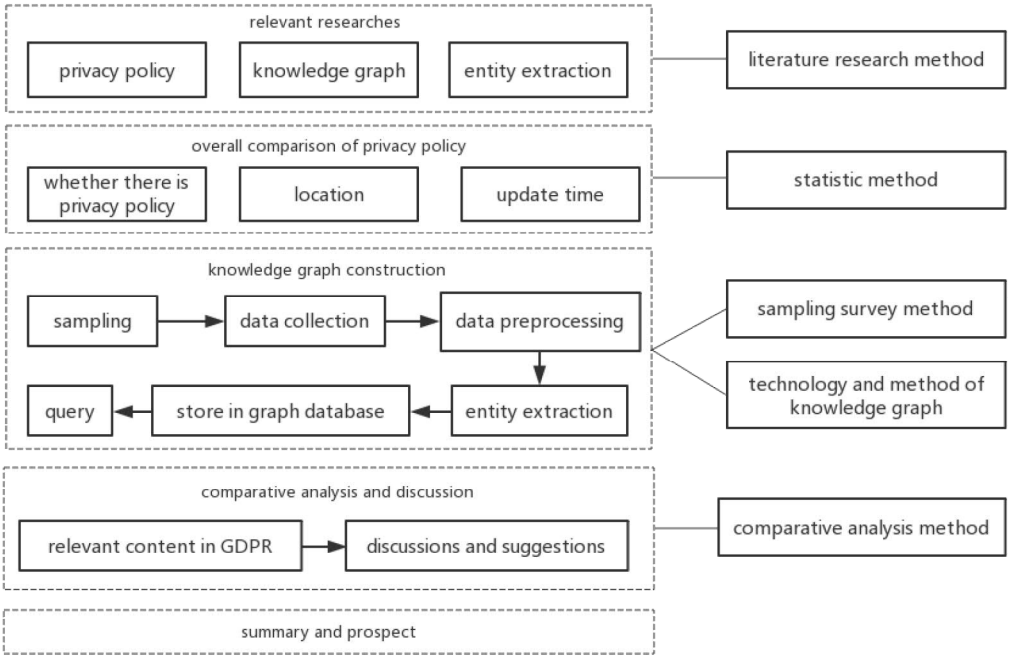


Figure 1-2 Structure of the study

This study consists of six chapters, which is as follows:

- Chapter 1 is mainly about the study background, motivation, research questions, research gap and innovations, as well as an overall study plan, which starts this study in a meaningful manner.
- In Chapter 2, researches of privacy policy, basic knowledge of knowledge graph and methods of entity extraction are introduced. The literature research method is used.
- Chapter 3 introduces the principle of Conditional Random Fields (CRF), the processes to achieve the study and the graph database, as well as the evaluation methods, aimed at laying a good foundation for the following empirical research.
- Chapter 4 is the empirical part. At first, samples are chosen and corpuses are collected. Then, overall comparisons are done from three characteristics of privacy policy, by statistic method. Next, “collected information” and “contact

us" of the privacy policy are cut out, for the construction of knowledge graphs. In this chapter, sampling survey and technologies of knowledge graph are used.

- Chapter 5 is to further discuss the comparative results with GDPR, as well as some suggestions according to the findings.
- Chapter 6 describes the contribution and limitation of this study, and the future work is expected.

2 THEORETICAL BACKGROUND

2.1 Privacy policy

2.1.1 Privacy and personal information

Whether "privacy" and "personal information" are different is controversial. Although most webmasters agree that there are differences between them, we can often see some privacy-related clauses, which are all about the collection, dissemination, sharing, use and publication of personal information.

In the environment of new media, privacy has gradually evolved from "the right of individual solitude" to "the freedom and right of individuals control over their own information" (Zhang, 2017). The information privacy right is the right of individuals, groups or institutions to decide when, in what way and to what extent their information is made known to others (Westin, 1967). The core of the information privacy right is personal information. Personal information does not refer specifically to sensitive, private or embarrassing personal information, but to the relationship between information and all its owners, that is, no matter the information is sensitive, as long as it is identifiable exclusively to individuals, it should be regarded as personal information (Kang, 1997). Therefore, in this study, we do not specially distinguish the usage of “privacy” and “personal information”.

2.1.2 Three modes to protect online privacy

About the protection mode of the information privacy right in cyberspace, Michelfelder (2001) pointed out that there are three solutions to protect online privacy: technical protection, self-discipline protection and legal protection. Table 2-1 clearly shows the three modes to protect online privacy.

Table 2-1 Three modes to protect online privacy

Technical protection	**Self-discipline protection**	**Legal protection**
Privacy protection system, such as Personal Data Valuts	Industry Self-Discipline Policy	Laws, regulations, like General Data Protection Regulation

The technical solution is mainly based on certain technological tools, which is chosen by Internet users themselves. Users can monitor and protect their online privacy by selecting online privacy protection software or systems. For example, the privacy

protection software - Personal Data Valuts can help users participate in data-sharing decisions, and protect online privacy by management of data policies (Mun et al., 2010); SPARCLE recognizes the elements of rules through parsers to automatically parse the organization's privacy policies (Brodie, Karat & Karat, 2006).

The self-discipline mode mainly protects online privacy by establishing industry associations and promulgating policies. Members participating associations must consciously abide by industry self-regulation policies. Some associations also issue logos of certification organizations to their members, such as TRUSTe Online Privacy Seal, California Company, USA, to inform netizens that the enterprise complies with industry self-regulation conventions (Benassi, 1999).

The legal mode is to establish the basic principles and specific legal provisions and systems of online privacy protection by making laws, and take corresponding judicial or administrative relief measures, such as GDPR.

2.1.3 Researches about privacy policy

Privacy policy is a new way to solve the problem of privacy protection. The protection based on privacy policy has become a new research field in security protection, and its researches and applications have attracted much attention (Liu, Wan & Li, 2016). The researches about privacy policy are mainly about the principle and content, as well as the practical application.

In the aspect of principle and content about privacy policy, the Organization for Economic Cooperation and Development (OECD) gives eight principles (*category, data, purpose, recipient, access, retention, disputes and remedies*) from the perspective of fair information (Organisation for Economic Co-operation and Development, 2002). Kwon (2010) divides the content of network privacy policy into four parts: *Categories, Purposes, Options and Retentions*.

The practical application researches about privacy policy are quite many. The W3C organization has put forward Platform for Privacy Preferences (P3P). By matching the privacy policy of the website and the user's privacy preference, the P3P decides whether to allow the user's privacy data to be used by the website (Marchiori, Cranor, Langheinrich, Presler-Marshall & Reagle, 2002). Pollach (2007) suggests that the privacy policy is not to build trust, but to clarify the issues. He thinks privacy policies can be improved through content, language and presentation format. Xu, Dinev, Smith & Hart (2011) made an empirical study of privacy policy on four different websites: e-commerce websites, social networking websites, financial websites and health care websites. They found that when users perceive more usefulness of privacy policy, their perceived privacy control increases significantly, and their perceived privacy risk

decreases, thus reducing privacy concerns. From the point of view of emotion and cognition of online users, Li, Sarathy & Xu (2011) proved that privacy policy negatively affects online users' perception of privacy risk, thereby improving users' disclosure intention. Tesfay et al. (2018) develop a system to help users to summarize the privacy policy, instructed by laws such as GDPR, based on machine learning and natural language processing techniques.

2.2 Knowledge graph

2.2.1 Predecessors of knowledge graph

Knowledge graph was put forward in 2012. Before that, it had many predecessors, as is shown in Figure 2-1.

In 1960, the Semantic Web was proposed as a method of knowledge representation, mainly for natural language understanding (Davies, Fensel & Van Harmelen, 2003). In the 1980s, the philosophical concept “Ontology” was introduced into the field of artificial intelligence to depict knowledge (Moens & Steedman, 1987). In 1989, Berners-Lee (1989) first proposed the "universal linked information system". It was the beginning of the invention of the World Wide Web. Then it evolved from hypertext links to semantic links in 1998. In 2006, Berners-Lee et al. (2006) highlighted the nature of the Semantic Web was to build links between open data. In 2012, Google released its search engine products based on knowledge graph (Steiner, Verborgh, Troncy, Gabarro, & Van de Walle, 2012).

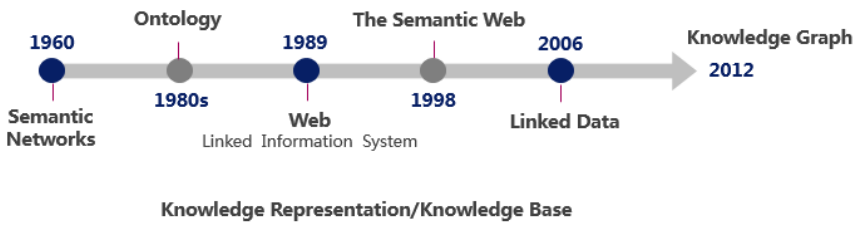


Figure 2-1 Predecessors of knowledge graph (based on Paulheim, 2017)

2.2.2 Types of knowledge graph

Knowledge graphs can be divided into open knowledge graphs and domain knowledge graphs according to the category of knowledge it contains (Pujara, Miao, Getoor, & Cohen, 2013; Deshpande et al., 2013).

The knowledge contained in the open knowledge graph is not divided into domains. It aims at acquiring all important concepts, entities, events and their relationships. Large-scale open knowledge graphs include: Google knowledge graph (Singhal, 2012), Satori (Ma, Crook, Sarikaya & Fosler-Lussier, 2015) and Probase (Wu, Li, Wang & Zhu, 2012) of Microsoft, YAGO (Suchanek et al., 2007) Series of Musk Plonk Institution, KnowItAll (Etzioni et al., 2004) of University of Washington, NELL (Carlson, Betteridge, Kisiel, Settles, Hruschka & Mitchell, 2010) of Carnegie Mellon University, DBpedia (Lehmann et al., 2015) and Freebase based (Bollacker, Evans, Paritosh, Sturge & Taylor, 2008) on crowd-sourcing, as well as WordNet (Miller, 1995) of Princeton University, see table 2-2.

Table 2-2 Examples of open knowledge graphs

Large-scale open knowledge graphs	**Development institution or method**
Google knowledge graph	Google
Satori and Probase	Microsoft
YAGO Series	Musk Plonk Institution
KnowItAll	University of Washington
NELL	Carnegie Mellon University
DBpedia and Freebase	Crowd-sourcing
WordNet	Princeton University

The knowledge contained in the domain knowledge graph is domain-specific. It is usually constructed to describe the knowledge in a particular domain. Domain knowledge graphs include: GeoName (GeoNames) of geographical domain, DBLife (DeRose et al., 2007) of academic domain, UniProKB (Bairoch et al., 2005) of biological domain and Linked Movie Database (Hassanzadeh & Consens, 2009) of movie domain, see table 2-3.

Table 2-3 Examples of domain knowledge graphs

Domain knowledge graphs	**Which domain**
GeoName	Geographical domain
DBLife	Academic domain
UniProKB	Biological domain
Linked Movie Database	Movie domain

Although open knowledge graphs have made great progress in recent years, such as YAGO, DBpedia, which have reached the scale of tens of millions of entities, the knowledge they contain is usually factual or conceptual knowledge, resulting in they are not well applied in the specific domain. Domain knowledge graphs are usually semantic networks constructed by extracting entities and relationships among entities from specific resources. The knowledge they contain is usually highly domain-specific. In many domains, high-quality domain knowledge graphs can often be embedded in the practical application.

2.2.3 The architecture of knowledge graph

The architecture of knowledge graph includes logical architecture and technical architecture (Qiao, Yang, Hong, Yao & Zhiguang, 2016).

In logical architecture, knowledge graph can be divided into schema layer and data layer. The schema layer is the conceptual model of knowledge graph. It defines the norms of the upper concepts of knowledge graph and typically constructed by ontology. The schema layer uses ontology to model domain concepts, so as to standardize and constrain the various factual expressions of the data layer. Therefore, the schema layer pays more attention to the relationship between concepts, the constraints between concepts and attributes, and the formal expression based on these concepts. The data layer is based on the schema layer, and stores knowledge by facts as units. It is a triple of "entity1-relation-entity2" or "entity-attribute-value". The data layer contains a large number of instances and their relationships. (Li, 2018) Figure 2-2 shows the logical architecture of knowledge graph.

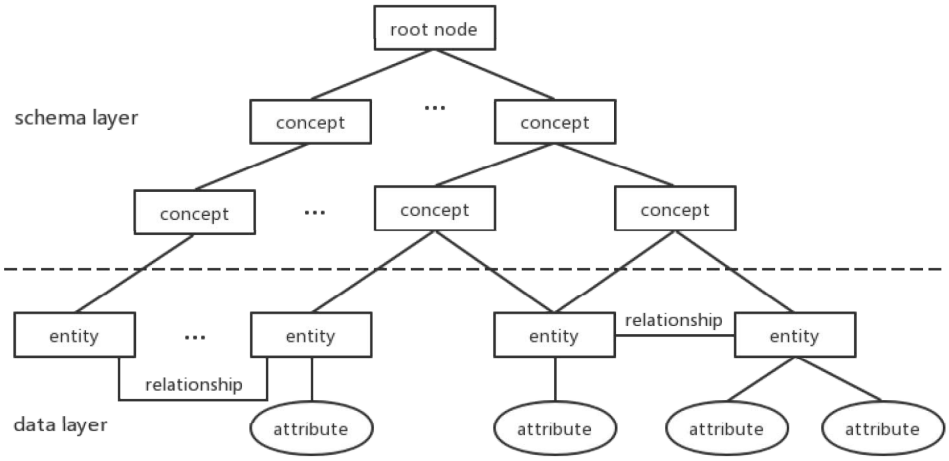


Figure 2-2 The logical architecture of knowledge graph (based on Li 2018, 7)

The technical architecture of knowledge graph is introduced from the angle of construction. Figure 2-3 gives the overall framework of knowledge graph technology, in which the part of dotted line is the process of the construction and updating of knowledge graphs (Qiao et al., 2016). In figure 2-3, knowledge can easily be extracted from structured data because of its high degree of standardization. Semi-structured and unstructured data are not structured enough to obtain knowledge directly, so it is necessary to extract entities associating with procedures such as entity extraction, relationship extraction and attribute extraction, and then store them in the knowledge base. (Liu, 2018)

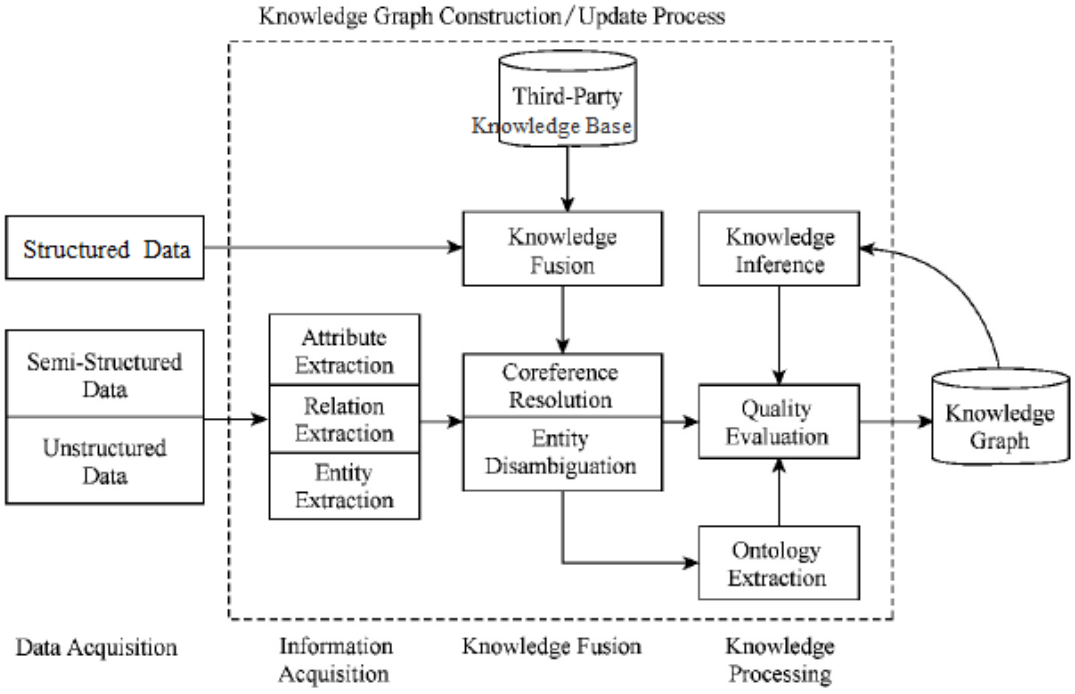


Figure 2-3 Technical architecture of knowledge graph (adopted from Qiao et al., 2016)

There are two ways to construct knowledge graph: top-down and bottom-up (Liu, 2018). Top-down construction method is based on ontology, using highly structured encyclopedias and other websites as data sources, extracting ontology and rule constraints and filling them into the knowledge base, while bottom-up construction method is to directly recognize entities, attributes and relationships from the open data like web data, and add them to the knowledge graph.

Using bottom-up approach to construct knowledge graph is an iterative updating process. Every iteration includes the following three steps (Qiao et al., 2016):

- Information acquisition. That is extracting entities/concepts, attributes and relationships from semi-structured and unstructured data sources.

- Knowledge fusion. Knowledge fusion is the fusion of entities which refer and represent the same meaning, from multiple sources. In the procedure, coreference resolution and entity disambiguation may be used.
- Knowledge processing. After the knowledge fusion, the quality of new knowledge needs to be evaluated according to the ontology constructed. Then the new knowledge is stored in the knowledge graph. Through knowledge inference, the knowledge is assessed again to ensure the quality of new knowledge. It is also an iterative procedure.

2.2.4 The construction method of knowledge graph

The construction method of knowledge graph can be devided into three situations according to data sources. Here we give four examples from four categories of data sources.

- Knowledge graph constructions based on Web-based encyclopedia resources. Gregorowicz and Kramer (2006) have successfully extracted more than two million concepts from Wikipedia, which can be graphed to more than three million terms.
- Knowledge graph constructions based on structured data. A knowledge graph can be regarded as a collection of triples such as entity-relationship-entity triples. Resource description framework (RDF) is developed by W3C (McBride, 2004). Its essence is a data model. It provides a unified standard for describing entities and resources. A RDF triple is a triple described by RDF. Tools such as D2R (Bizer, 2003) can convert traditional relational data which is structured data into RDF triples. Based on the triples, a knowledge graph is constructed.
- Knowledge graph constructions based on semi-structured data. Shinzato and Torisawa (2004) proposed a method of automatically extracting the upper and lower relationships from HTML documents to assist the construction of knowledge graph, in order to overcome the shortcoming of the traditional methods of making specific language model to obtain the upper and lower relationships.
- Knowledge graph constructions based on unstructured data. KnowItAll (Etzioni et al., 2004) and NELL (Carlson et al., 2010) use incremental iteration method to learn high quality triples from a large amount of web data.

At present, knowledge graph has been successfully applied in intelligent question answering (Hixon, Clark & Hajishirzi, 2015; Yahya, Barbosa, Berberich, Wang & Weikum, 2016), semantic search (Su et al., 2015; Dalton, Dietz & Allan, 2014),

recommendation system (Sigurbjörnsson & Van Zwol , 2008), text understanding (Wang, Zhang, Feng & Chen, 2014; Hakkani-Tür, Celikyilmaz, Heck, Tur, & Zweig , 2014), entity disambiguation (Cucerzan, 2007), machine translation (Wang et al., 2014) and other application scenarios. These applications help to process knowledge quickly and accurately.

2.3 Entity extraction

2.3.1 The process of enttity extraction

Entity extraction, also known as Named Entity Recognition (NER), refers to the automatic recognition of person name, place name, institution name and other named entities from text datasets (Chinchor & Robinson, 1997). Entity extraction is the core and basic part of the knowledge graph. Its precision and recall rate directly affect the quality of knowledge graph.

Named entity recognition, as a basic research work in the field of Natural Language Processing (NLP), has been applied to many aspects of NLP (Nadeau & Sekine, 2007). Especially with the continuous development of information extraction technology, NER has become a hot research topic.

At present, many new methods have been applied to NER. Great breakthroughs have been made. The recognition of person name, place name, organization name, time expression, numerical expression has been in a high quality in the open field. With the deepening of the research on NER, entity recognition technology has also been greatly developed.

The general process of NER is shown in Figure 2-4.

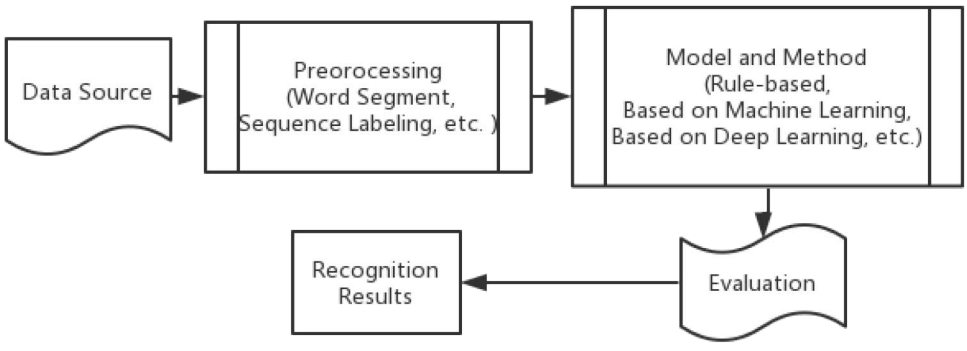


Figure 2-4 General process of NER (based on Shao, 2017)

2.3.2 Supervised method

Supervised method regards NER as a sequence-labeling problem. Sequence labeling models include Hidden Markov Models (HMM) (Bikel, Miller, Schwartz & Weischedel, 1998), Maximum Entropy Markov Models (MEMM) (McCallum, Freitag & Pereira, 2000) and Conditional Random Fields (CRF) (Lafferty, McCallum & Pereira, 2001). These models are all based on a large number of tagged corpus, define a series of entities, and obtain feature-based discriminant rules through learning.

Hidden Markov Model considers context information. The solution obtained in the test is the global optimum solution, and the optimal Markov chain is obtained. This is not possible by traditional classification algorithm (Bikel et al., 1998). The shortcoming of hidden Markov model is that it assumes that observable variables are independent, and that the constraints on observable variables are the words themselves, which limits the selection of features. For example, features such as word number, document frequency, location, etc., which are very predictive to the entity type, can not be used conveniently.

Maximum Entropy Markov Model only calculates the probability of hidden variables under a given observable variable, and transforms the Hidden Markov Model into a discriminant model (McCallum et al., 2000). It overcomes the shortcomings of the Hidden Markov Model and makes it easy to use various features. However, it also brings a new problem of label bias.

Conditional Random Fields model transforms the conditional probability in the Maximum Entropy Markov Model into the eigenfunction form. Through training, the weights of different features are obtained. Viterbi algorithm is usually used to solve the problem in testing (McCallum & Li, 2003). The Conditional Random Fields model overcomes the label bias problem of the Maximum Entropy Markov Model and have a good effet on entity extraction, but it also has the problem of slow class training. In this thesis, we tend to choose Conditional Random Fields model as entity extraction method becuase we can label a number of sequence manually.

2.3.3 Semi-supervised method

Semi-supervision is also called weak-supervision. The idea of semi-supervised machine learning was established in the process of researching extracting structured data from texts to construct biological knowledge base (Craven & Kumlien, 1999).

The main technology is Bootstrapping, which provides only a few labeled data for initial learning. For example, a system for identifying disease names needs users' examples. The system searches for sentences containing disease names and identifies

their context. Then the system searches for other disease names in the context that is identical to the previous examples. The process of learning is a continuous cycle of above, discovering new contexts, discovering new disease names, generating a large number of basic disease names and contexts.

The methods of identifying contextual environment include Collins and Singer (1999) adopt template method and Cucchiarelli and Velardi (2001) adopt parsing tree. Semi-supervised method can achieve good results under the condition of a small amount of labeled dataset and a large number of unlabeled dataset.

2.3.4 None-supervised method

The most typical method of unsupervised learning is clustering. For example, different named entities are brought together in a similar context.

There are other unsupervised methods, including the method of Transfer Learning (Alfonseca & Manandhar, 2002), the method based on mutual information (Etzioni et al., 2005). The method based on external resources is when there is no corpus for a particular domain, we can use external resources such as WordNet (Miller, 1995) to have transfer learning. The method based on mutual information is to classify the given words and determine the type of input.

2.4 GDPR (General Data Protection Regulation)

General Data Protection Regulation (GDPR) sets relatively high and strict data protection standards, including chapters of principles, rights of data subjects, controller and processer and so on, totaling eleven chapters. Privacy policy is a statement that shows how data controllers, usually companies, collect and use the users' data. We can understand privacy policy as a statement how data controllers process their users' data, and GDPR as the standard to constraint the data controllers processing users' data. Furthermore, there are five stages of contract agreement and contact, in which the fourth stage is "policy comparison" between "service consumer" and "trusted privacy service" or "service provider" and "trusted privacy service" (Allison et al., 2009). According to their model, in our study, the discussions with GDPR implements the fourth stage between "service provider" and "trusted privacy service". Therefore, this chapter is going to discuss the comparative results generated in chapter 4, according to GDPR. Before that, several definitions adopted from GDPR are given as follows (General Data Protection Regulation, Chapter 1 Article 4):

- ***"personal data"*** *means any information relating to an identified or identifiable natural person (**"data subject"**); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person;*
- ***"processing"*** *means any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction;*
- ***"controller"*** *means the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data; where the purposes and means of such processing are determined by Union or Member State law, the controller or the specific criteria for its nomination may be provided for by Union or Member State law.*

These definitions will show frequently in the discussion part. Informally, in this thesis, "data subject" refers to the user, "controller" refers to the Internet company who collect users' personal data, and "personal data" refers to the data collected from users.

3 METHODOLOGY

3.1 Study design

To have the comparative study of Chinese and European Internet companies' privacy policy, the following steps are supposed to be considered:

- The selection of sample Internet companies and the preparation of the privacy policy corpus.
- Having an overall comparative study by simple statistic method based on the sample data.
- Preprocessing of the original privacy policy corpus - word segmentation and part-of-speech tagging (POS tagging) by POS tagger (A toolkit for POS tagging), format edition by UltraEdit (A text editor).
- Having the entity extraction by CRF++ (A toolkit for entity extraction).
- Constructing the knowledge graphs by Neo4j (A graph database).
- Getting information from the knowledge graphs by Cypher.

Figure 3-1 shows the process of the comparative study. The bracketed content is how to realize the steps or with what tools.

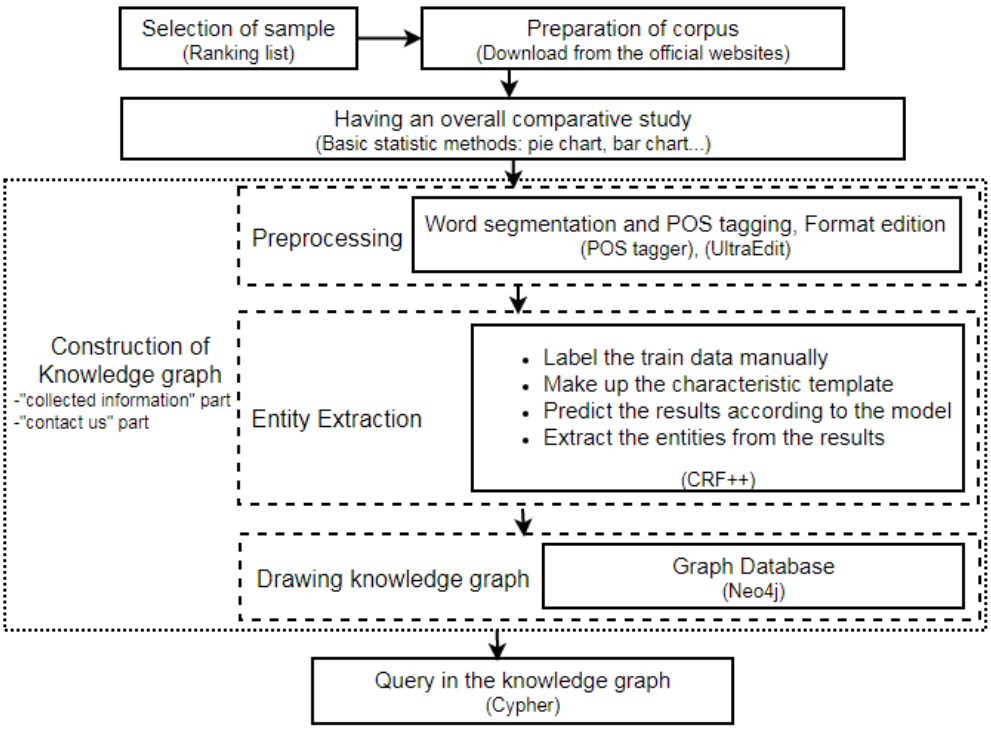


Figure 3-1 The process of the comparative study

The first two steps are relatively easy to achieve. We can select sample Internet companies by ranking lists and prepare the privacy policy corpus by downloading from the companies' websites. The step of having an overall comparative study is easy to achieve. For example, we can count the mouse click times reaching the privacy policy for comparative study. Therefore, the following sections are mainly about the construction of knowledge graph and its query language.

3.2 Word segmentation and POS tagging

Word segmentation is cutting a string of written language into its component words (Gambell & Yang, 2006). POS tagging is a crucial part in Natural Language Processing (NLP). POS tagging is labeling the part of speech of each component word which is processed by word segmentation. These two steps are preprocess for entity extraction.

POS tagger (Tsuruoka, 2005) is a toolkit developed in University of Tokyo. Tsuruoka and Tsujii (2005) proposed a bidirectional inference algorithm for sequence labeling problems such as POS tagging. Based on this, they developed a toolkit named POS tagger which offers fast tagging (2400 tokens/sec) with a state-of-the-art accuracy (97.10% on the WSJ corpus). The tagger uses an extension of Maximum Entropy Markov Models (MEMM), in which tags are determined in the easiest-first strategy. Table 3-1 shows the POS tags and their meanings.

Table 3-1 POS tags and their meanings (adopted from Gole, 2015)

Tag	Description
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun

Tag	Description
PRP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	to
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Whdeterminer
WP	Whpronoun
WP$	Possessive whpronoun
WRB	Whadverb

UltraEdit is a powerful text editor, which can edit text, hexadecimal and ASCII codes. It can edit multiple files at the same time. Even if it opens large files, it will not slow down. After word segmentation and POS tagging, we can use UntraEdit to edit the format of the corpus to make it meet the input requirements of CRF++.

3.3 The method and realization of entity extraction

3.3.1 CRF (Conditional Random Fields)

In section 2.3.2, we have introduced the three supervised methods - Hidden Markov Models (HMM), Maximum Entropy Markov Models (MEMM) and Conditional Random Fields (CRF), for entity extraction. CRF is mainly constructing model for the target sequence according to the observation sequence (Lafferty et al., 2001). CRF combines the characteristics of production models. It not only avoids the strong independence hypothesis of HMM, but also effectively solves the label bias problem in MEMM, and has a good effect in sequence labeling. The most commonly used linear chain structure of CRF is shown in figure 3-2.

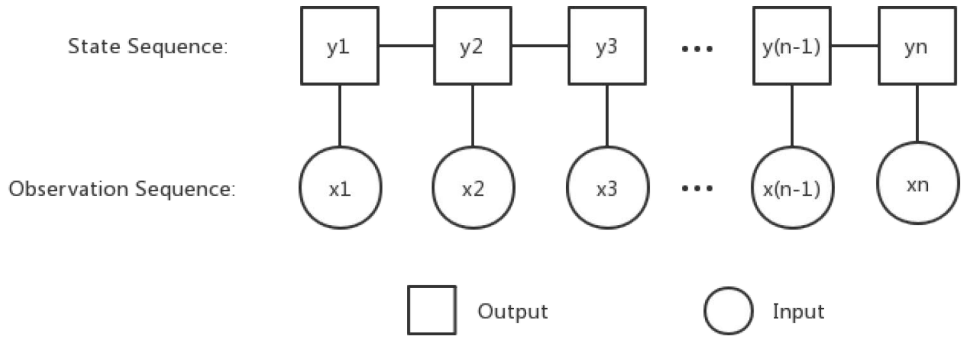


Figure 3-2 Liner chain structure of CRF (based on Lafferty et al., 2001)

According to CRF, the entity recognition process of text corpus is as the sequence labeling process of text corpus. That is, treating every sentence in the text as an observation sequence, and regrading every word in the observation sequence as a symbol. Giving every symbol a label, for observation sequence X=(x1, x2, x3, ..., xn) and state sequence Y=(y1, y2, y3, …, yn), P(Y|X) is the conditional probability distribution of the state sequence Y under the known condition X. According to CRF, there is:

$$P(Y|X) = \frac{1}{z(x)} \exp(\sum i \sum k \lambda_k f_k(y_{i-1}, y_i, x) + \sum i \sum k u_k g_k(y_i, x)) \quad (1)$$

In equation (2), fk(yi-1, y1, x) and gk(yi, x) are both characteristic function, in which fk(yi-1, y1, x) represents whether there is the "k" characteristic of the state sequence in the "i" position of current input, and it depends on the current state of y1 as well as the previous state of yi-1. λ_k and u_k is the weight value of the corresponding characteristic function. z(x) is all state sequences' planning factors:

$$\mathrm{Z(x)} = \sum y \exp(\sum i \sum k \, \lambda_k f_k(y_{i-1}, y_i, x) + \sum i \sum k \, u_k g_k(y_i, x)) \quad (2)$$

Giving an input sequence "X" of Internet companies' privacy policy texts, the target is to find the most possible labeling result sequence "Y", that is:

$$y^* = argmax\mathrm{P(Y|X)} \quad (3)$$

All above equations are based on the proposed literature of CRF (Lafferty et al., 2001).

3.3.2 CRF++

To realize the core step of the construction of knowledge graph, we choose CRF++ toolkit to have the entity extraction. CRF++ is a simple, customizable, and open source implementation of Conditional Random Fields for segmenting/labeling sequential data. CRF++ is designed for generic purpose and will be applied to a variety of NLP tasks, such as Named Entity Recognition, Information Extraction and Text Chunking (CRF++: Yet Another CRF toolkit).

Using CRF++ to extract entities from the text, the main idea is using train dataset to construct model in order to predict test dataset. First, the corpus need to be divided into two parts — train dataset and test dataset. We label the train dataset by hand for training the test dataset, while the test dataset is the target dataset where you want to recognize entities. To make sure the quality of the results, the train dataset and test dataset had better come from the same source, because the toolkit predicts the test dataset according to the train dataset. Then we need to label characteristics for the train dataset and edit its format for the training by CRF++. After training, it will generate a model file, and we can use this model to predict labeling the entities in the test dataset.

The CRF++ is strict with the text format of input train dataset and test dataset. If the format is not conformed to the standard, CRF++ cannot run. The format is as follows:

- The train and test files are presented in multiple lines, using a token to represent one line. Each token contains a number of columns consisting of observation words and a space or a tabular form separates the characteristics. The last column of characteristic represents the label of the observed value.
- A sentence usually consists of several tokens, and each sentence is distinguished by a blank line.

- The last column of token is the correct labeling form for training.

Taking one of the CRF++ self-contained examples named “basenp” as example, figure 3-3 is the example format CRF++ requires. In figure 3-3, every token in the corpus is divided into three columns, which are word self, part of speech and BIO label (Chao et al., 2007), separately. In BIO labeling, “B” means the beginning of the phrase, “I” means the following part of the phrase and “O” means it is not the phrase we want (Getting Started In: Sequence Labeling).

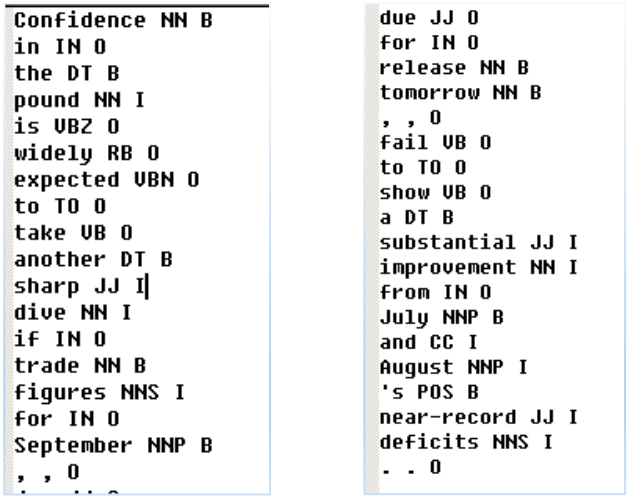


Figure 3-3 An example which shows the format CRF++ requires

In CRF++, the characteristic template is used to describe the characteristics of train dataset and test dataset. In the experiment, we need to construct characteristic template for calculating the model. There are two characteristic templates — Unigram template and Bigram template. For Unigram template, when a template is given, CRF++ automatically generates a set of feature functions (funcl…funcN) for each input category. The total number of characteristic functions generated is L*N, where L is the number of categories and N is the number of characteristics extended by the template. For Bigram template, the system automatically combines the current output token with the last token output, and the total number of characteristics generated is L*L*N (CRF++: Yet Another CRF toolkit). Figure 3-4 shows an example of the characteristic template.

```
# Unigram
U00:%x[-2,0]
U01:%x[-1,0]
U02:%x[0,0]
U03:%x[1,0]
U04:%x[2,0]
U05:%x[-1,0]/%x[0,0]
U06:%x[0,0]/%x[1,0]

U10:%x[-2,1]
U11:%x[-1,1]
U12:%x[0,1]
U13:%x[1,1]
U14:%x[2,1]
U15:%x[-2,1]/%x[-1,1]
U16:%x[-1,1]/%x[0,1]
U17:%x[0,1]/%x[1,1]
U18:%x[1,1]/%x[2,1]

U20:%x[-2,1]/%x[-1,1]/%x[0,1]
U21:%x[-1,1]/%x[0,1]/%x[1,1]
U22:%x[0,1]/%x[1,1]/%x[2,1]

U23:%x[0,1]

# Bigram
B
```

Figure 3-4 An example characteristic template

In figure 3-4, the format of each row % x [row, col] represents a sub-template, where the parameter “row” represents the deviation from the current token (relative position), while the parameter “col” represents the absolute position of the column, and the starting values of both parameters are zero.

The execution process of CRF++ is in the Windows system’s DOS commend. To execute the program, the train and test dataset, as well as the execution files must be put in the same file. The basic train and test instructions are as figure 3-5. The result is stored in output.txt file.

```
crf_learn template train.data model
crf_test -m model test.data >> output.txt
```

Figure 3-5 Train and test instructions

If we want to evaluate the results, the evaluation program — conlleval.pl can help us evaluate the precision rate, recall rate and F measure. Figure 3-6 is the instruction.

```
perl conlleval.pl <output.txt >result.txt
```

Figure 3-6 Evaluation instructions

This section records the experimental steps of training, testing and evaluation of entity recognition. According to the entity recognition labeling results, we can easily extract the entities we want. For example, we can copy the labeling result in Microsoft Excel and use formulation function to realize the extraction of entities.

3.3.3 Evaluation indicators

Precision and recall rate are commonly used as evaluation indicators in the field of entity extraction. Precision and recall rate are two contradictory evaluation indicators. Generally speaking, the higher the precision is, the lower the recall rate is; on the contrary, the higher the recall rate is, the lower the precision is. Therefore, the entity extraction performance is usually evaluated by the F-score, which is a comprehensive weighted indicator of the precision rate and recall rate. The calculation equations for precision, recall and F measure are as follows (Leacock, Chodorow, Gamon & Tetreault, 2010):

$$\mathrm{P} = \frac{TP}{TP+FP} = \frac{Number\ of\ Entities\ Correctly\ Recognized}{Number\ of\ Entities\ Recognized} * 100\% \quad (4)$$

$$\mathrm{R} = \frac{TP}{TP+FN} = \frac{Number\ of\ Entities\ Correctly\ Recognized}{Number\ of\ Entities\ in\ the\ Text} * 100\% \quad (5)$$

$$\mathrm{F} = \frac{2*P*R}{P+R} * 100\% \quad (6)$$

In above equations, P denotes the correct rate of entity recognition, R denotes the ability to recall entities, TP (True Positives) denotes the number of entities correctly identified, FP (False Positives) denotes the number of entities incorrectly identified, and FN (False Negatives) denotes the number of entities incorrectly identified as non-entities.

3.4 Neo4j graph database

A graph database management system, referred to as a graph database, is an online database management system with create, read, update and delete etc. methods that expose a graph data model (Robinson, Webber, & Eifrem, 2013). Neo4j is a NoSQL graph database management system. It inherits the advantages of graph database: good performance, flexibility and agility. Neo4j graph database has the following four basic features (Chen, 2017):

- Nodes, relationships and attributes are the three basic elements of a graph database.

- The attributes of nodes and relationships are a collection of key-values.
- Each relationship has a start node and an end node connected to each other.
- In most cases, attributes may not be required.

Cypher is an expressive (yet compact) graph database query language. It is specific to Neo4j until now, but it does not prevent it becoming a very concise and easy-to-understand graph database query language. Common-used Cypher clauses is as follows (Robinson et al., 2013):

- MATCH. MATCH is usually used to match the data in the database to obtain the data satisfying the query conditions.
- WHERE. WHERE is not a clause in the strict sense. It is generally used as part of the MATCH clause to specify the conditions that the query needs to meet. This is similar to WHERE in SQL.
- RETURN. RETURN specifies which queries need to return.
- CREATE. CREATE can be used to create nodes, relationships, attributes, etc.

4 COMPARISON PROCESS

4.1 Privacy policy corpus preparation

As is mentioned in the section 1.3.1, Internet companies is a good choice to study privacy policy. To have the comparative study, at first we need to choose sample Internet companies of China and Europe separately. The idea to find ranking lists. Then we go to the official websites to find and download their privacy policies.

4.1.1 Privacy policy corpus of Chinese Internet companies

On July 27, 2018, the China Internet Association and the Information Center of the Ministry of Industry and Information Technology jointly released the list of the top 100 Internet companies in China in 2018 (XinhuaNet, 2018). At first we plan to choose the top 20 Internet companies as privacy policy corpus. However, there are two companies—"2345.com" and "XinhuaNet" for which we cannot find privacy policies in their official websites. The solution is abandoning these two companies and postponing in the ranking list until we find 20 eligible Internet companies. Appendix 1 shows the privacy policy links of Chinese Internet companies we collected. The retrieved time is February 24th, 2019.

Normally, the privacy policy is at the bottom of the official website, but some Internet companies such as Tencent and Baidu set specific privacy protection platforms. Their privacy policies are in the privacy protection platforms. The 20 privacy policy texts are downloaded for the following comparative study.

4.1.2 Privacy policy corpus of European Internet companies

There is not an official ranking list of European Internet companies. We used a website named Informilo that has selected "the 25 hottest Internet companies in Europe" (Informilo). The top 20 Internet companies are chosen and their privacy policy texts are downloaded successfully. Appendix 2 shows the privacy policy links of European Internet companies we collected. The retrieval time was February 24th, 2019.

Since this ranking list is not new, two companies — Shazam and Fotolia have been acquired by Apple and Adobe, using their patent companies' privacy policies. This does not affect our study, because the parent companies' privacy policy are also worth

studying. The 20 privacy policy links in appendix 2 are all at the bottom of the official websites.

4.2 Overall comparative analysis

4.2.1 Whether there is privacy policy link in the official website

As is mentioned in section 4.1.1, the original plan is downloading the privacy policies of top 20 Chinese Internet companies in the ranking list. Nevertheless, we cannot find the privacy policy links in two companies' official websites. While all privacy policy links of top 20 European Internet companies in the ranking list were found. Using pie charts, figure 4-1 shows in the top 20 ranking list of Chinese and European Internet companies, whether there is privacy policy links in the official websites.

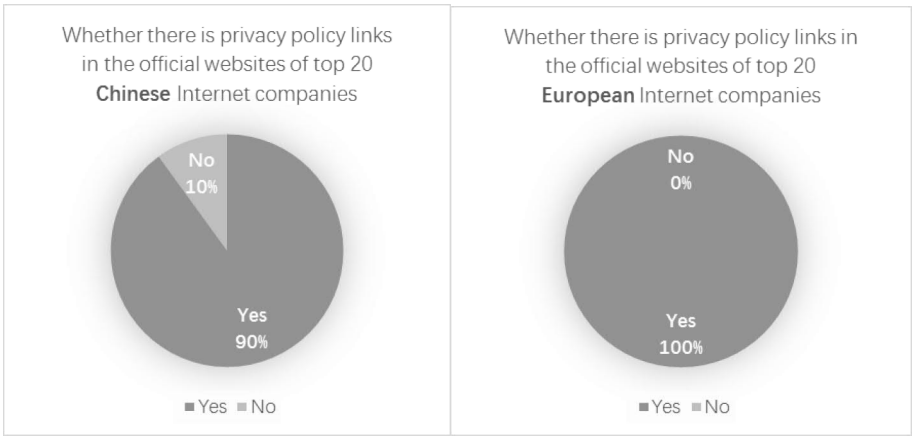


Figure 4-1 Comparison — whether there is privacy policy link in the official website

Although the sample is small — only 20, we can find some Chinese Internet companies lack awareness of setting privacy policies in their official websites. Moreover, when we postpone in the ranking list for another two Chinese Internet companies for their privacy policies, the 21st company does not have privacy policy on its official website, either. In some sense, it can reflect that the awareness of setting privacy policies of Chinese Internet companies is not as good as that of European Internet companies.

4.2.2 The click times to reach the privacy policy

To some extent, the location of the privacy policy links in the official website reflects the company's attention of users' privacy protection. Appendix 3 and 4 separately show the steps and mouse click times to reach the privacy policy of sample Chinese and European Internet companies.

To have a clear view to analyze the click times to reach the privacy policy of Chinese and European Internet companies, a bar chart is created, see figure 4-2.

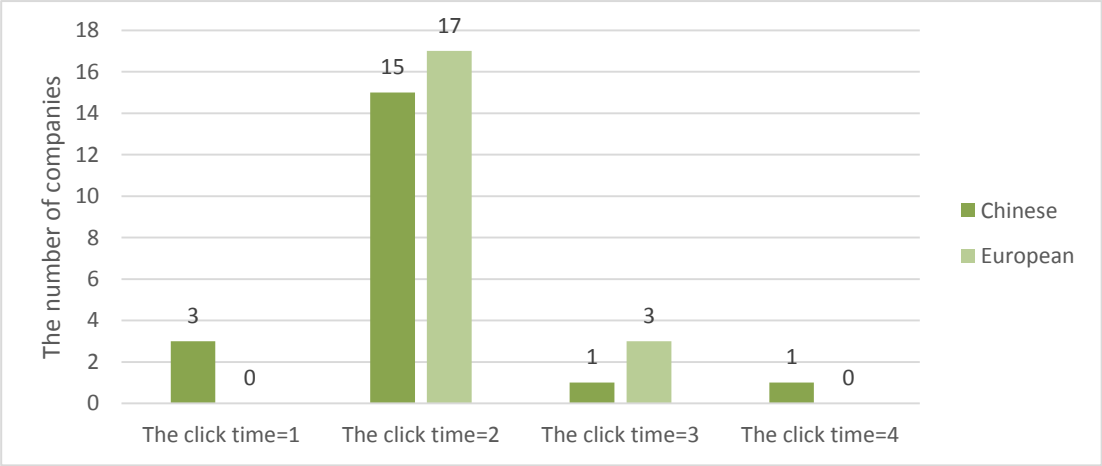


Figure 4-2 The click times to reach the privacy policy

From an overall perspective to analyze figure 4-2, the number of Internet companies whose click time=2 is the highest, at 15 and 17. Most Internet companies put the privacy policy links in the homepage of their official websites, no matter Chinese or European Internet companies. It can reflect that through many years of appealing of protecting information privacy, the Internet companies have realized the importance of protecting users' privacy by setting privacy policies and show them on the obvious location in their official websites.

In the angle of comparison, European Internet companies are more united than Chinese Internet companies. All sample European Internet companies put the privacy policy links at the bottom of their official websites, so two steps is easy to find the privacy policy. There are just three European Internet companies — Layar, Shazam and Klarna, setting one more hyperlink to classify the information in details. This result is similar as Irene Pollach's. He studies of 49 websites in four categories: retail websites, news websites, travel websites and portal websites, and finds that 90% of websites only need one click on the main page to reach privacy policies (Pollach, 2006). However, the gap among Chinese Internet companies in privacy policy setting is a little large. In the original sample, two companies' privacy policies cannot be found. Additionally, three companies set specific privacy policy platforms, but for another two companies — Sohu and Kingsoft — privacy policies are hard to find.

4.2.3 The update time of privacy policy

The update time in a way can also reflect the company's attention of users' privacy protection. If a company does not update its privacy policy for a long time, we can speculate this company does not pay much attention on users' privacy protection in cyberspace. The update time is ofen showed at the beginning and the end of the privacy policy text, so we count the update time without striking a blow. The line chart, see figure 4-3, shows the change trend of the privacy policy update time of the sample companies.

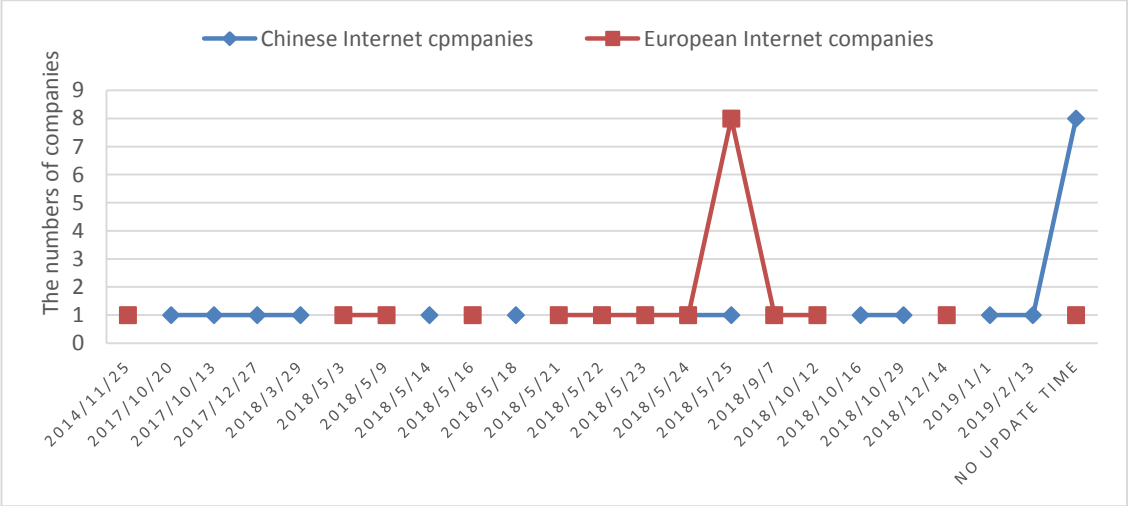


Figure 4-3 The update time of privacy policy

It can be seen from the figure 4-3 that the privacy policy update time of European Internet companies concentrates in May, especially on May 25th, 2018, which is the date of the GDPR come into force. It reflects that GDPR brought some influences for privacy policies of European Internet companies. Most companies updated their privacy policy in May, before or after the date of acting GDPR. It is noticeable that there is a company – Layar — whose updating time of privacy policy is November 25th, 2014. Layar was acquired by the UK company Blippar in June 2014 but Layar did not inherit its parent company's privacy policy and stopped updating the privacy policy since November 25th, 2014. It is reported that one of Layar's founder want to buy back Layar (Palmer, 2019), but this cannot fully explain why Layar does not update its privacy policy. We can regard it as a special case.

While many Chinese Internet companies do not label their update time, which is not friendly for the readers who want to know the when the privacy policy start to be

effective. The updating time of Chinese Internet companies is separate. It seems GDPR has little influence for Chinese Internet companies on the privacy policy update time.

4.3 What knowledge graphs shall be constructed

4.3.1 Problem interpretation

This section is to solve the research question — "What kind of knowledge graphs shall be constructed". In chapter 3, we know that the core of constructing knowledge graphs in this study is entities extraction. So the research question "What kind of knowledge graphs shall be constructed" can be changed into "What kind of entities shall we extracted from the privacy policy text". Whatever Chinese or European privacy policy, their structures are similar. Basically, the privacy policy consists of "What information we collect", "How we use that information", "Your rights", "Use of Cookies/Beacons", "Contact us" and other parts (Shen, 2017), though their expressions are slightly different. For the two parts — "What information we collect" and "Contact us", entities we want to collect are clear, which will be introduced in details in the following. While the other parts — "How we use that information", and "Use of Cookies/Beacons" are too complicated to extract specific entities, waiting for future research. Therefore, in this study, "Collected information" and "Contact us" are chosen to construct knowledge graphs separately.

4.3.2 "Collected information" of the privacy policy

The data collected by Internet companies from users is important for user's privacy protection. It is the foundation of online privacy protection. Once collected information is controlled abusively, the problem of online privacy protection becomes serious. Pollach (2007) lists five categories of questions which are the key privacy concerns among Internet users. One of the category is data collection. Furthermore, Attaran and VanLaar (1999) introduce "how to secure your personal information and company data". One of the tips they give is checking what personal information the companies would collect. They think "the best policy is one that does not collect any of your personal information and the worst is one that does not tell users that information is being collected or how it will be used". According this think, we can extract "collected information" as entities for constructing knowledge graphs and see if the companies would totally tell what personal they collect.

In this study, we tend to extract the personal data collected by companies such as name, email address, as entities for constructing knowledge graphs, to see what kind of personal information would Chinese and European Internet companies collect, and have comparisons and analysis.

4.3.3 "Contact us" of the privacy policy

When users want to look for help upon their online privacy, the effective contact ways are significant. Pollach (2006) count the contact methods in privacy policies of fifty well-known websites. In his research, contact information includes postal address, phone number, email address and email form and two companies do not show their contact information in their privacy policies. Allison et al. (2009) put forward five stages of contract agreement and contact stages. The third stage is "privacy inquiry" between "service provider" and "service consumer". To realize the "privacy inquiry", an effective contact method shown on the privacy policy is important.

Generally, the contact data can be divided into postal, telecom and online data (Karjoth & Schunter, 2002). In the construction of privacy policy knowledge graphs of "contact us" part of this study, contact methods — email, postal address, phone number, online service and reply time would be extracted as entities separately.

4.4 Privacy policy knowledge graphs — collected information

4.4.1 Text preprocessing

In the section 4.1, the data source is prepared. The privacy policy corpus has been stored in the style of text. Here we take the part of "What information we collect", or be called "collected information", as the corpus for constructing knowledge graphs.

First, we cut out the part of "What information we collect" from the 40 intact sample privacy policy texts. Then we delete the special format to meet the requirement of the toolkit of word segmentation and part-of-speech tagging (POS tagging).

POS tagger can be operated in Windows System's DOS commend window. Figure 4-4 shows the operation and execution of POS tagger.

```
C:\Users\apple>cd C:\Users\apple\Desktop\core of thethesis\part-of-speech tagger
\University of Tokyo\postagger

C:\Users\apple\Desktop\core of thethesis\part-of-speech tagger\University of Tok
yo\postagger>tagger<1.txt>>output1.txt
loading ./models/model.bidir.0
loading ./models/model.bidir.1
loading ./models/model.bidir.2
loading ./models/model.bidir.3
loading ./models/model.bidir.4
loading ./models/model.bidir.5
loading ./models/model.bidir.6
loading ./models/model.bidir.7
loading ./models/model.bidir.8
loading ./models/model.bidir.9
loading ./models/model.bidir.10
loading ./models/model.bidir.11
loading ./models/model.bidir.12
loading ./models/model.bidir.13
loading ./models/model.bidir.14
loading ./models/model.bidir.15

C:\Users\apple\Desktop\core of thethesis\part-of-speech tagger\University of Tok
yo\postagger>
```

Figure 4-4 Word segmentation and POS tagging by POS tagger

The result is output in text. In the output text, the word and its label are separated by "/", which does not conform to the prescribed format of the CRF++. Here we use UltraEdit to replace "/" into tabular forms. The corpus after preprocessing is shown in figure 4-5.

```
,       ,
common  JJ
name    NN
,       ,
first   RB
names   NNS
,       ,
gender  NN
,       ,
date    NN
of      IN
birth   NN
,       ,
sponsor NN
and     CC
or      CC
sponsorees      NNS
,       ,
delivery        NN
address NN
```

```
Data    NNP
automatically   RB
collected       VBN
by      IN
us      PRP
When    WRB
using   VBG
Gameforge       NN
services        NNS
,       ,
system  NN
and     CC
user-related    JJ
data    NNS
is      VBZ
collected       VBN
automatically   RB
and     CC
without IN
any     DT
further RBR
```

Figure 4-5 The corpus after preprocessing

4.4.2 Entity extraction

To have entity extraction by CRF++ toolkit, we need to divide the corpus into two parts — train dataset and test dataset. Typically, the proportion of train dataset and test dataset is from 50% - 50% to 90% -10% (Larose & Larose, 2014). Here we choose 50% -50% for train dataset and test dataset. The entity extraction of Chinese and European corpus is handled separately. We randomly choose 10 pieces from the 20 pieces Chinese Internet companies corpus as train dataset, and the rest 10 pieces is test dataset. The operation of European Internet companies corpus is the same. Appendix 5 shows the details of the distribution of train and test datasets.

For the train dataset, we label the entities we want to extract on the third column. Tabular forms separate the third-column's label and the second-column's label as well. Because the corpus we use is the part of "What information we collect" in the intact privacy policy, we want to extract the entities which can reflect what information the companies collecting from users, such as date of birth, name, location and so on. Figure 4-6 is a part of train dataset.

```
you             PRP     O
to              TO      O
provide         VB      O
information             NN      O
related         VBN     O
to              TO      O
personal                JJ      B
identity                NN      I
,               ,               O
such            JJ      O
as              IN      O
personal                JJ      O
identification  NN      O
including               VBG     O
ID              NNP     B
card            NN      I
,               ,               O
passport                NN      B
,               ,               O
and             CC      O
driver          NN      B
license         NN      I
```

Figure 4-6 A part of train dataset

Now the train dataset and test dataset are prepared, so the CRF++ toolkit can be used to recognize the entities we want. Following the instructions of the figure 4-7 and operating on Windows system's DOS commend, we can get the prediction results of the test dataset. Figure 4-7 shows the operation and execution of CRF++.

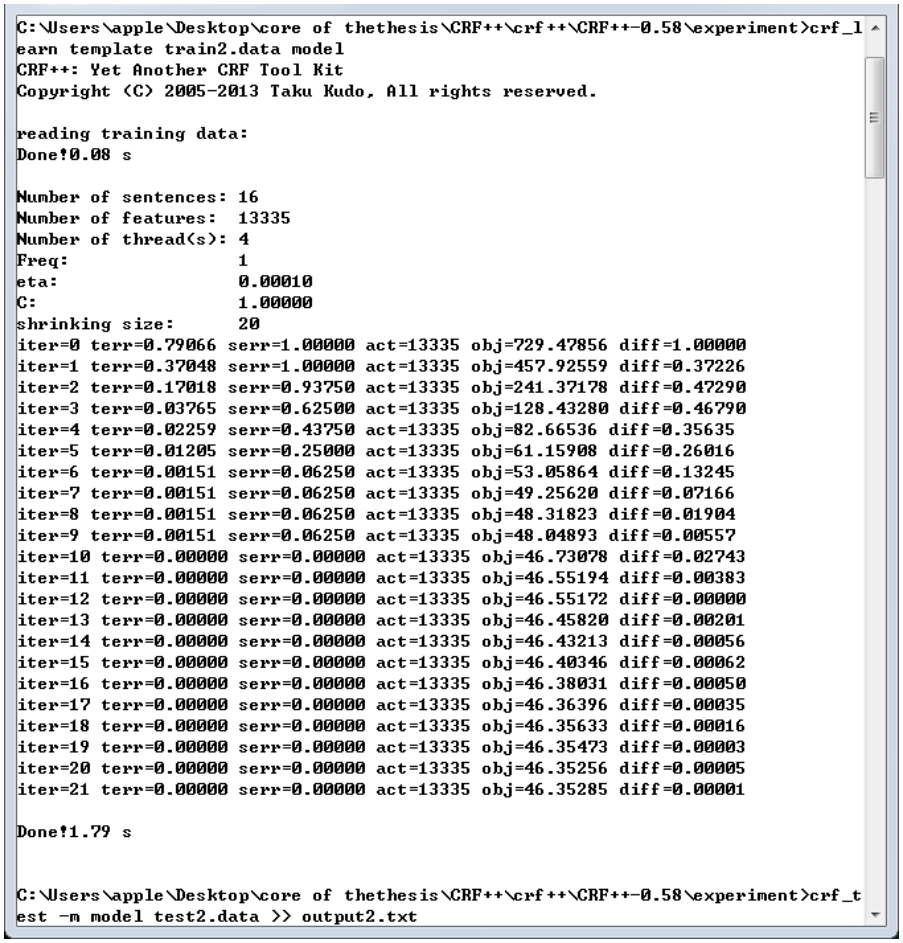


Figure 4-7 Entity Recognition by CRF++ toolkit

The CRF++ recognizes 376 entities in Chinese Internet companies' privacy policy corpus, and 512 entities in European Internet companies' privacy policy corpus. From the results of the evaluation program — conlleval.pl, the precision rate is 68.83%, the recall rate is 80.83%, and F-score is 74.35.

According to the prediction label of CRF++, we can easily extract the entities. The entities are corresponding to their companies, so the relationship between them is "contain", that is, the company "contain" these entities.

4.4.3 Draw knowledge graph by Neo4j

Neo4j supports a variety of data import and storage methods. It can be imported manually, that is, the creation of nodes, attributes and relationships can be achieved by writing Cypher statements one by one, as well as supports batch import of csv files. It also supports importing data from mainstream relational databases. Because there are many nodes in this study, we use the Cypher statement to read and import the data from the csv file.

To draw knowledge graph, we need to put the entities we extracted and the corresponding company names together in the text file, making them entity pair, and separate them with commas. Then we need to convert the text file into csv file, and use the following Cypher statements (see figure 4-8) to import the data from the csv file.

```
LOAD CSV WITH HEADERS FROM "file:///experiment.csv" AS row
with row
merge(c:Company{name:row.Company})
merge(d:CollectedData{name:row.CollectedData})
merge(c)-[:contain]->(d)
```

Figure 4-8 Importing data in Cypher

After importing data into Neo4j, we can use Cypher to query and display some part or the panorama of privacy policy knowledge graphs we need. Figure 4-9 and figure 4-10 is the panoramas of “CollectedData” entities of Chinese and European Internet companies’ privacy policy, separately.

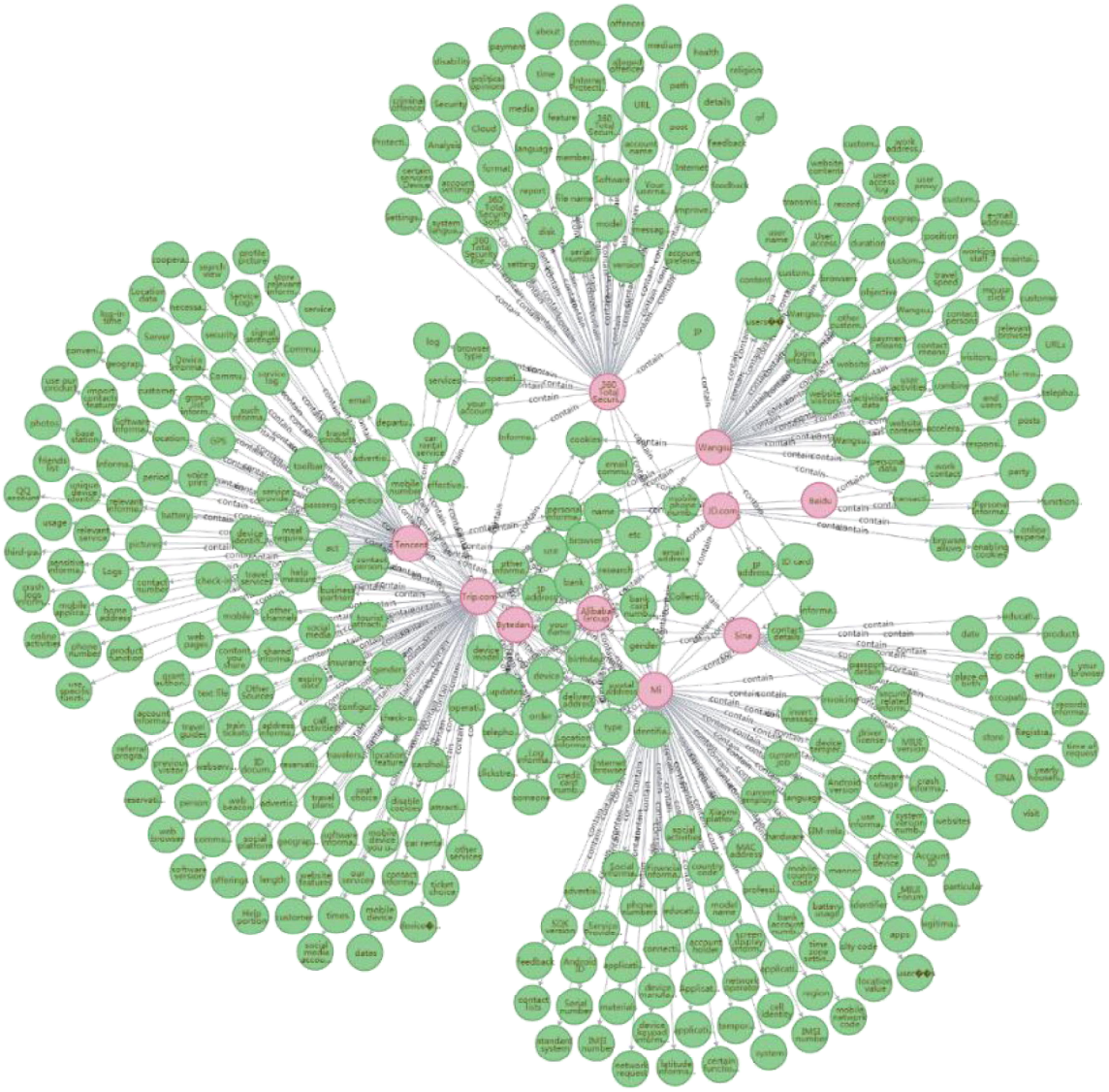


Figure 4-9 Privacy policy knowledge graph of Chinese Internet companies — collected information

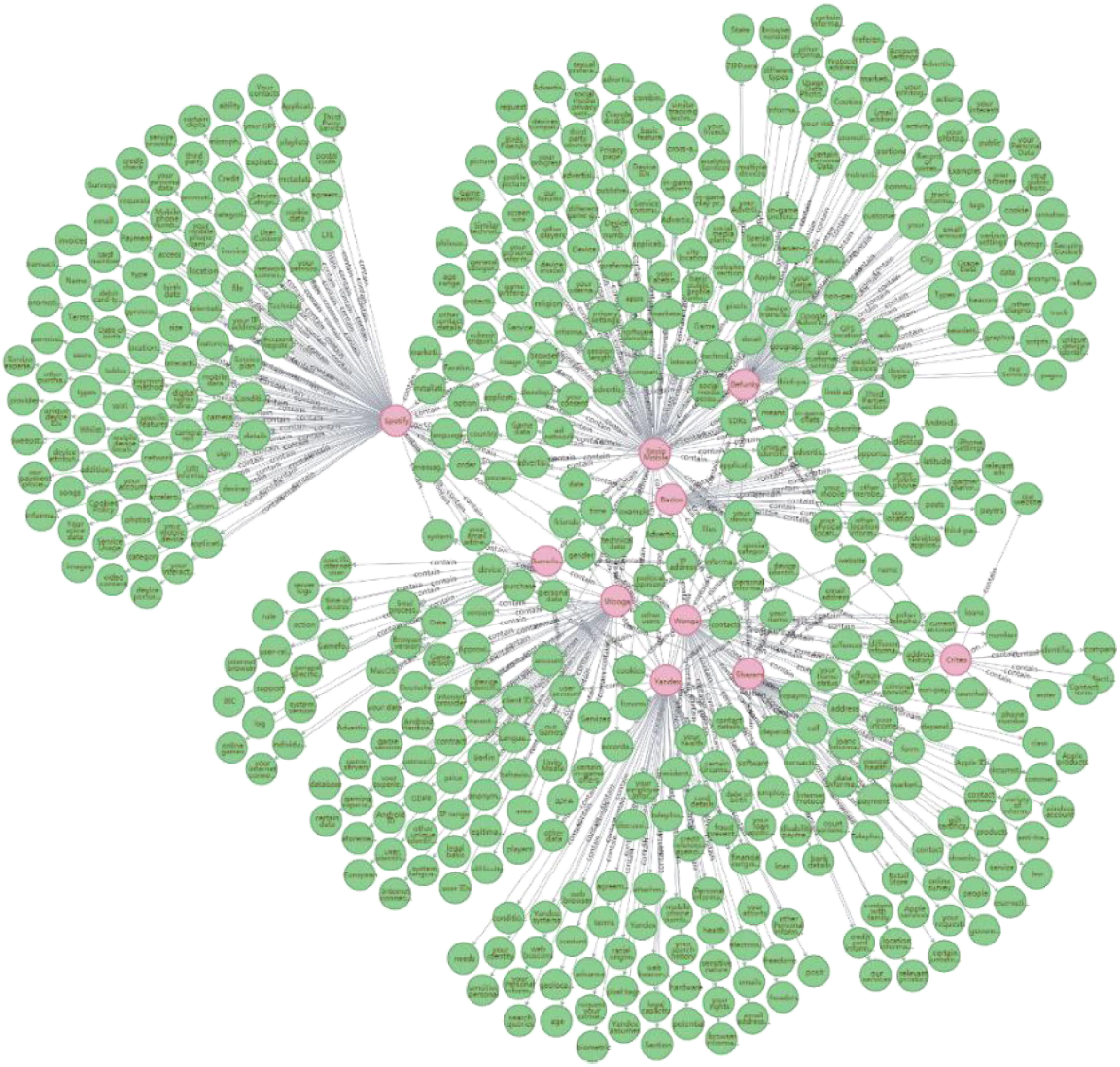


Figure 4-10 Privacy policy knowledge graph of European Internet companies — collected information

From the panorama of the two knowledge graphs of figure 4-9 and figure 4-10, no matter Chinese or European Internet companies, the entities of every company is relatively in a large number, some even at 125 entities — Rovio Mobile. The phenomena of companies containing the same entities is easy to observe. The differences of the two knowledge graphs are obvious. In the “collected information” of privacy policy, entities extracting from European companies are more than those extracting from Chinese companies. The number of entities extracting from European and Chinese companies are 512 and 376, separately. Furthermore, the network of entities of European companies is more complicated than that of Chinese companies. The relationships of entities of European and Chinese companies are 594 and 439, separately.

4.4.4 Query in the knowledge graphs

Among many entities extracting from the “collected information” part of privacy policy, we want to know which ones the companies contain in common. Since there are ten companies in each knowledge graph, the number of relationships orienting to one entity is natural number from 1 to 10. We start to try from the maximum number 10 with the following Cypher query statement, see figure 4-11.

```
MATCH (m:Company)-[r:contain]->(n:CollectedData)
With n, COUNT (r) AS x
WHERE x=10
return n
```

Figure 4-11 Query the entities shared by 10 companies

We first execute this statement in the privacy policy knowledge graphs of European Internet companies, but no result is found. Then we try 9, 8, 7, 6... until 2 with the query statement, and record every result see appendix 6. From appendix 6, we find some entities shared by European companies exist overlap in the meaning, such as “your name” and “name”, so we simplify the entities by deleting the similar entities in the smaller x value, manually. Besides, some entities like “information”, “time”, and “service” are too general, we also abandon them. Table 4-1 shows the streamlined results of appendix 6.

Table 4-1 Simplified entities shared by European companies

The number of companies (x)	**Entities shared by x companies**
7	IP address
5	your name, your device
4	friends, gender
3	advertising, your consent, email address, address, cookies
2	messages, language, browser type, purchase, country, order, option, Facebook, interest, phone number, device identifiers, forums, your health, account, contact details, application, technical data, political opinions

From table 4-1, we can find “IP address” is the most popular entity that shared by seven companies in the “collected information” part of their privacy policy, followed by “your name”, “your device” and then “friends”, “gender”. These entities are all basic

and important personal information. We choose entities shared by five or more companies as objects and draw knowledge graph, see figure 4-12.

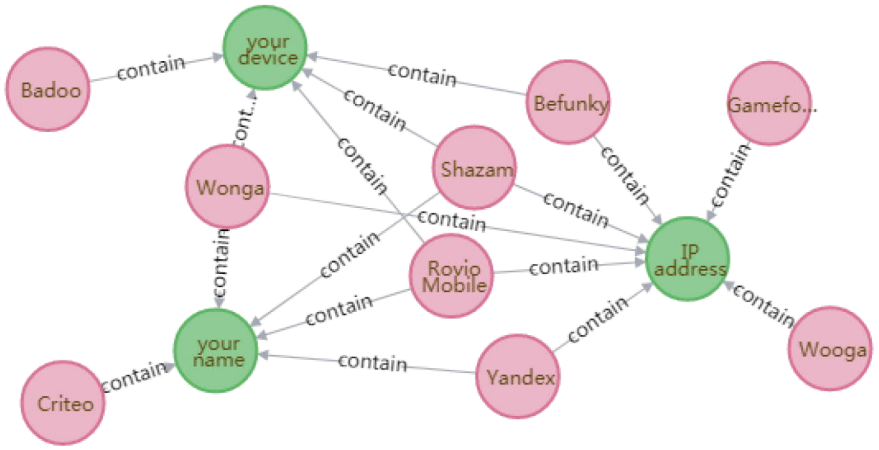


Figure 4-12 Entities shared by five or more than five European Internet companies

Figure 4-12 gives top three words that European Internet companies most likely to collect from users. They are “IP address”, “your device” and “your name”.

The same process is done in the privacy policy graph database of Chinese Internet companies. Appendix 7 shows the entities shared by Chinese companies. Table 4-2 is the simplified version of appendix 7.

Table 4-2 Simplified entities shared by Chinese companies

The number of companies (x)	**Entities shared by x companies**
6	your name, email address
5	other information
3	IP address, gender
2	clickstream data, operating system you use, postal address, telephone number, identifiable information, updates, your account, log, browser type, email communications, contact details, ID card, Location information, order, birthday, credit card number, cookies

We can see from table 4-2 that “your name” and “email address” are contained by six companies, followed by “other information”, and then “IP address” and “gender”. Except “other information”, the other entities is basic and important personal information, which is similar with the European situation. By reading the privacy policy text, the context of “other information” is usually “name, address and other information”. We find Chinese Internet companies prefer to use “and other information” to generalize the “collected information” rather than list them in details.

By Cypher query statement, figure 4-13 is generated to display the entities shared by five or above five Chinese Internet companies. There are also top three words — “your name”, “email address” and “other information”.

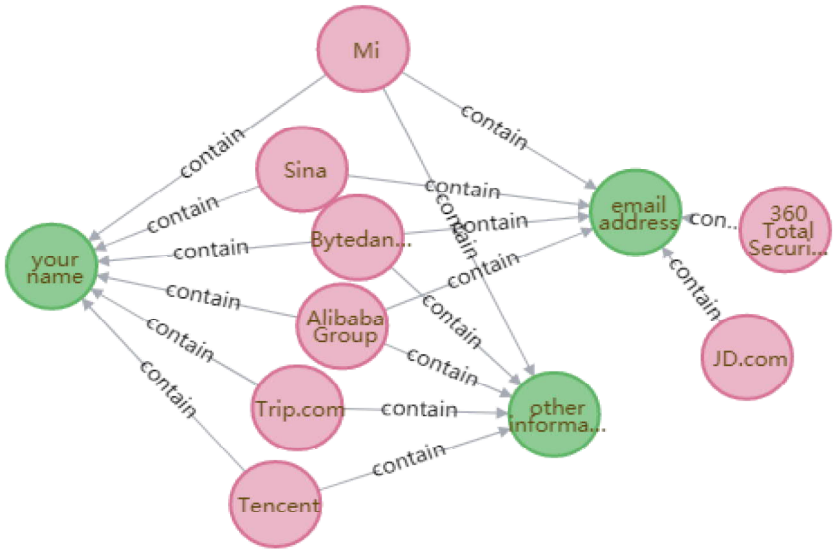


Figure 4-13 Entities shared by five or more than five Chinese Internet companies

The common entity between figure 4-12 and figure 4-13 is “your name”, which reflects users’ name is the one of the most popular word that Internet companies collect, no matter in China or in Europe.

To compare other differences between “collected information” of Chinese and European Internet companies, table 4-3 is made by deleting the common part of table 4-1 and table 4-2.

Table 4-3 Different entities between Chinese and European companies

Entities of Chinese companies	**Entities of European companies**
other information, clickstream data, identifiable information, updates, log, email communications, ID card, Location information, birthday, credit card number	friends, advertising, your consent, messages, language, purchase, country, option, Facebook, interest, device identifiers, forums, your health, technical data, political opinions

From table 4-3, we find that Chinese Internet companies collect some sensitive information like “ID card”, “location information”, and “credit card number”, but European Internet companies would not. They collect information such as “your health”, “Facebook”, “political opinions”. As is known to all, western people care more about health, most of them have Facebook and their political participatory is stronger compared with Chinese.

4.5 Privacy policy knowledge graphs — contact us

4.5.1 Privacy policy knowledge graph — email

In this section, the "contact us" parts are cut out from the intact privacy policy texts. The train dataset and test dataset are the same distribution as above, see appendix 5. The entities extraction of emails, postal addresses, phone numbers and reply time from the "contact us" part of privacy is the same as the process mentioned above, so we just give the results rather than the processes in this section. Additionally, there is another one method provided by companies to contact with them, which is online service. However, because the characteristics of the online service is not obvious - some are website links and some are hyperlinks named "click here", we will count the it manually.

After preprocessing of the original corpus, we label the email entities in the train dataset according to BIO and then process it by CRF++. The email entities are shown in figure 4-14.

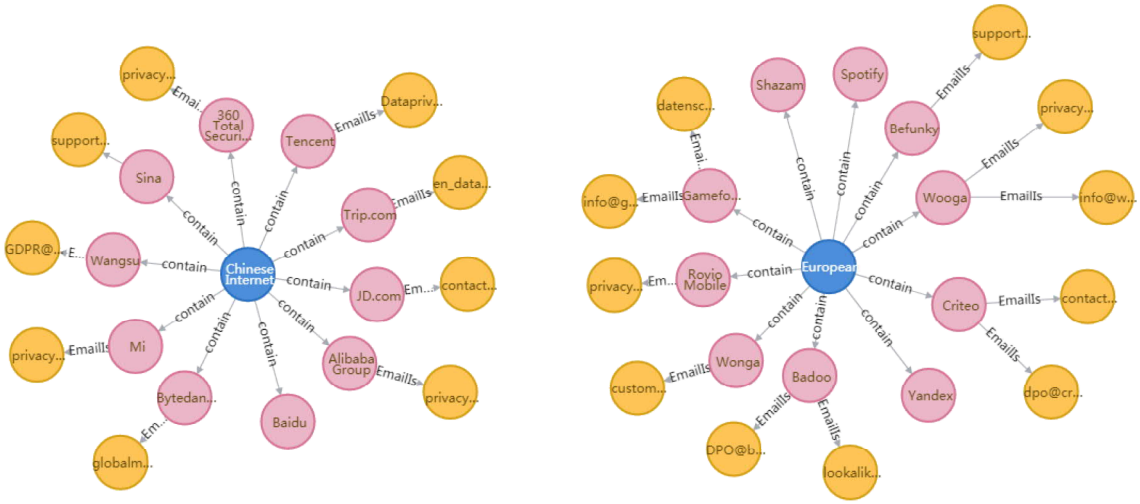


Figure 4-14 Privacy policy knowledge graph — email extraction

It can be seen from figure 4-14 that most Internet companies give emails in their privacy policies. In Chinese Internet companies, "Baidu" does not put its email contact information in its privacy policy. The other Chinese Internet companies show one email in the privacy policies. There are three European Internet companies — "Shazam", "Spotify" and "Yandex" cannot be found emails as contact information in their privacy policies, and three companies — "Gameforge", "Befunky" and "Criteo" have two email addresses in the privacy policies. The companies that give more than one email addresses usually want to subdivide the users' requirements for convenient management.

4.5.2 Privacy policy knowledge graph — postal address

The postal address entities are extracted, as is shown in figure 4-15.

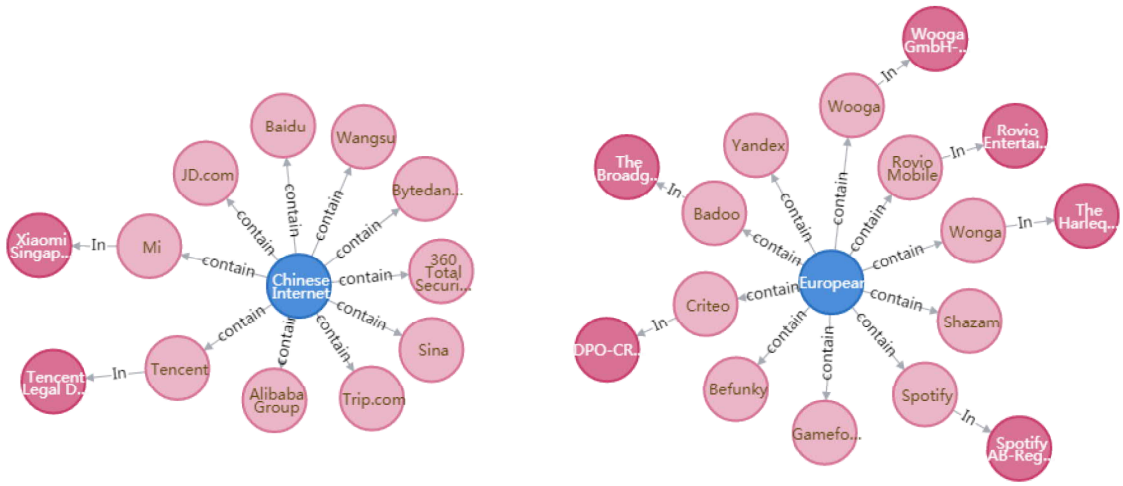


Figure 4-15 Privacy policy knowledge graph — postal address extraction

We can see from figure 4-15 that most Chinese Internet companies do not give their addresses in their privacy policies. While more than a half of European Internet companies put their detailed addresses in the privacy policies.

4.5.3 Privacy policy knowledge graph — phone number

The results of the phone number extraction are not good because the toolkit recognizes some postal code, so we delete these confuters by hand. Figure 4-16 shows the situation of the phone number in the privacy policies.

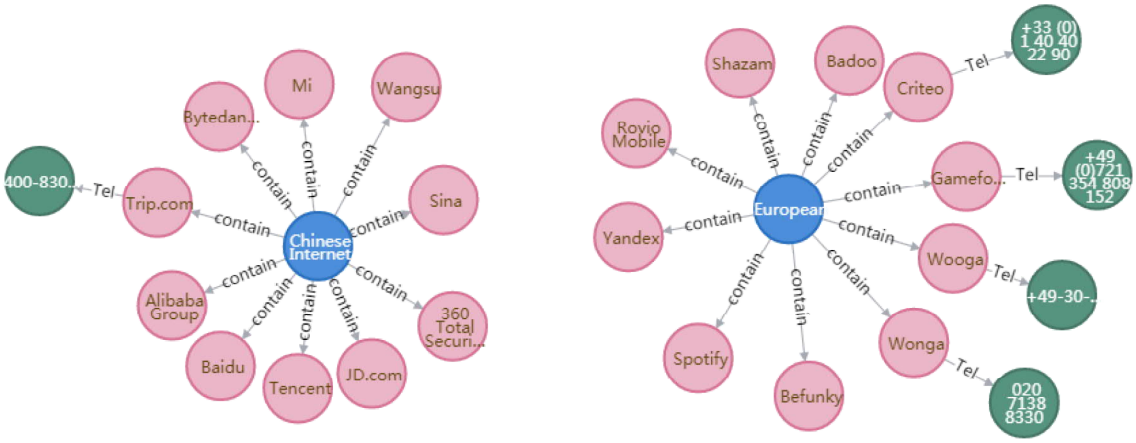


Figure 4-16 Privacy policy knowledge graph — phone number extraction

In the figure 4-16, most Internet companies do not offer the phone numbers in the privacy policies. Only one Chinese company — "Trip.com" provides the phone number as one of the contact methods. Four European companies also give their phone numbers.

4.5.4 Whether there is "online service" as the contact way

Table 4-4 shows whether there is "online service" for users to complain of these Internet companies.

Table 4-4 Whether there is "online service" as contact way

	Whether there is "online service" of Chinese Internet companies		**Whether there is "online service" of European Internet companies**	
1	Alibaba Group	Yes	Gameforge	No
2	Tencent	Yes	Spotify	Yes
3	Baidu	Yes	Befunky	Yes
4	JD.com	Yes	Shazam	No
5	Sina	No	Wonga	No
6	360 Total Security	Yes	Rovio Mobile	No
7	Mi	No	Badoo	No
8	Bytedance	No	Criteo	Yes
9	Wangsu	No	Wooga	No
10	Trip.com	Yes	Yandex	Yes

From table 4-4 we can count the proportion of Internet company setting "online service" in their privacy policies is 60% and 40% in China and Europe separately. But because the sample is too small, we just conclude that no matter for Chinese or European Internet companies, some companies prefer setting "online service" in their privacy policies, but some do not.

4.5.5 Privacy policy knowledge graph — reply time

In the "contact us" part of the privacy policy, some companies tell users the reply time to ensure the users' requests replying in a given time. We extract the reply time from the contact us part of the privacy policy, see figure 4-17.

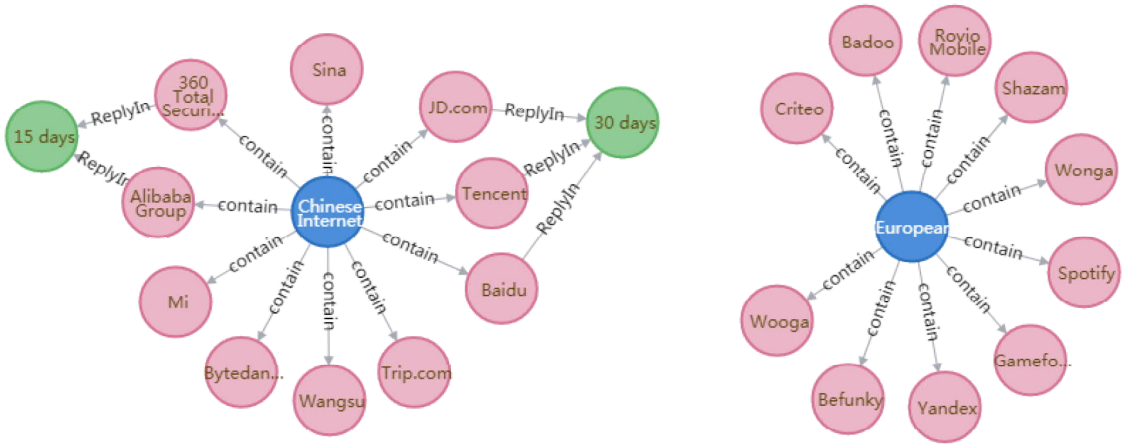


Figure 4-17 Privacy policy knowledge graph — reply time extraction

It is obvious that in the 10 sample European companies, no one gives the specific reply period in their privacy policies. For Chinese Internet companies, three companies write will reply in 30 days, and two companies express they will give feedback in 15 days.

4.6 Conclusions of the comparison results

In chapter 4, we compare the Chinese and European Internet companies' privacy policies from the overall perspectives by statistical methods, and from the privacy policy knowledge graphs of "connected information" and "contact us" part. The comparison results are concluded as follows:

- Some Chinese Internet companies do not set privacy policy link on their official websites.
- The locations of European Internet companies' privacy policies are united and obvious, at the bottom of the official websites. Some Chinese Internet companies' privacy policies are hard to find.
- The update time of the privacy policy of European Internet companies concentrates on May 25th, 2018, which is the date of the GDPR come into force. The update time of the privacy policy of Chinese Internet companies are relatively separately. Many Chinese Internet companies do not show the update time on privacy policies.
- From the privacy policy knowledge graphs of "collected information" part, entities extracting from European companies are more than those extracting from Chinese companies. The network of entities of European companies is more complicated than that of Chinese companies.

- In both Chinese and European Internet companies' privacy policies, users' name, email address and other basic personal information are collected. European Internet companies prefer to collect users' health data, Facebook account and othe personal information, while Chinese Internet companies would rather collect personal information such as location and credit card number.
- No mater Chinese or European Internet companies, they give at least one contact method in their privacy policies, in which "email" is the most popular contact way.
- Compared with European Internet companies, Chinese Internet companies do not prefer to provide "postal address" and "phone number" as the contact method.
- No sample European Internet companies offers "reply time" on their privacy policies. Half of the sample Chinese Internet companies offer the specific "reply time".

5 DISCUSSIONS AND SUGGESTIONS

5.1 Discussions of the comparative results with GDPR

5.1.1 Discussions of the overall comparative results

According to the Chapter 3 Article 12 Point 1 of GDPR (General Data Protection Regulation):

> *1. The controller shall take appropriate measures to provide any information referred to in Articles 13 and 14 and any communication under Articles 15 to 22 and 34 relating to processing to the data subject in a concise, transparent, intelligible and easily accessible form, using clear and plain language, in particular for any information addressed specifically to a child. The information shall be provided in writing, or by other means, including, where appropriate, by electronic means.*

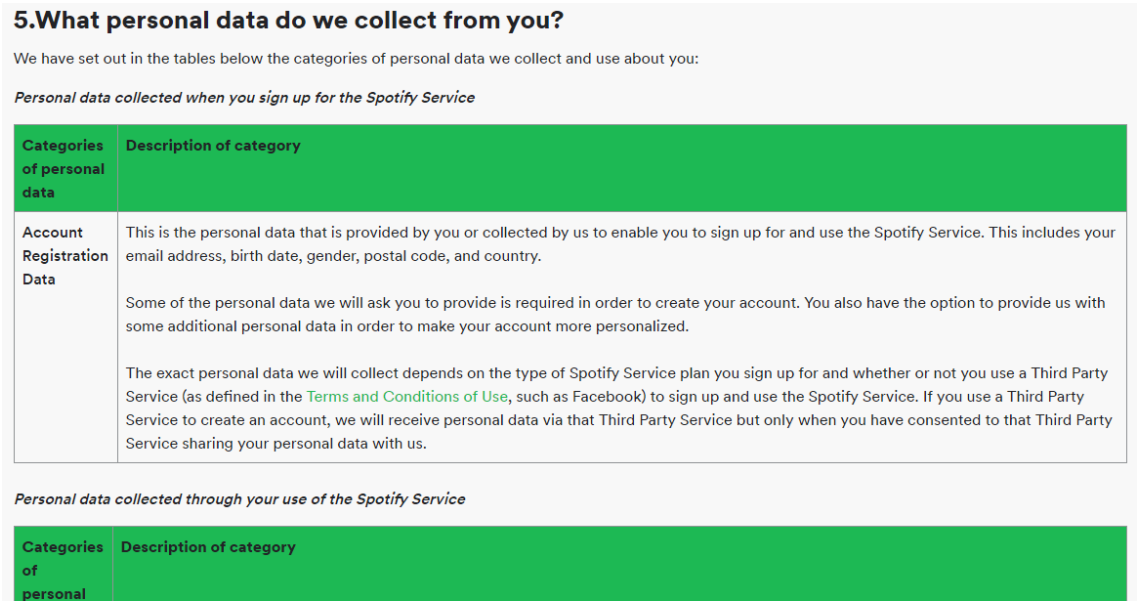
Privacy policy is one of the good measures for the data controller to take responsibilities because it is "concise, transparent, intelligible and easily accessible". However, we cannot find two Chinese Internet companies' privacy policy links in their official websites.

The mouse click times present the "easily accessible form" in the regulation in some sense. From figure 4-2, we notice that Chinese companies like "Sohu" set privacy policy links in a deep location, which is hard for users to find. This is against GDPR.

5.1.2 Discussions of the knowledge graph results — collected information

One result in section 4.4.3 is that in the "collected information" of privacy policy, entities extracted from European companies are more than those extracted from Chinese companies. The number of entities extracted from European and Chinese companies are 512 and 376, separately. One possible reason is the European companies' privacy policy is more detailed. By reading the "collected information" part of the sample privacy policy, the assumption is verified. For example, many European Internet companies list the collected data in a table and explain it in details, see figure 5-1, which conforms the "intelligible" requirement in GDPR, while few Chinese Internet companies do this. Another difference is the network of entities of European companies is more complicated than that of Chinese companies. Connected with the original text corpus,

we find the expression of European Internet companies' privacy policy is more united than that of Chinese companies. The "united" here means the European Internet companies prefer using the same word to express one content.



5.What personal data do we collect from you?

We have set out in the tables below the categories of personal data we collect and use about you:

Personal data collected when you sign up for the Spotify Service

Categories of personal data	Description of category
Account Registration Data	This is the personal data that is provided by you or collected by us to enable you to sign up for and use the Spotify Service. This includes your email address, birth date, gender, postal code, and country. Some of the personal data we will ask you to provide is required in order to create your account. You also have the option to provide us with some additional personal data in order to make your account more personalized. The exact personal data we will collect depends on the type of Spotify Service plan you sign up for and whether or not you use a Third Party Service (as defined in the Terms and Conditions of Use, such as Facebook) to sign up and use the Spotify Service. If you use a Third Party Service to create an account, we will receive personal data via that Third Party Service but only when you have consented to that Third Party Service sharing your personal data with us.

Personal data collected through your use of the Spotify Service

Categories of personal	Description of category

Figure 5-1 Spotify Company lists the collected personal data in tables

According to the Chapter 2 Article 5 Point 1 (a) of GDPR (General Data Protection Regulation):

1. Personal data shall be:
(a) processed lawfully, fairly and in a transparent manner in relation to the data subject ('lawfulness, fairness and transparency');

'Lawfulness, fairness and transparency' principle requires the personal data shall be processed lawfully, fairly and in a transparent manner. The data subjects have rights to know what their personal data would be collected by the data controllers. The Internet companies shall tell users what personal data they would collect as detailed as possible, which can reflect the "lawfulness, fairness and transparency".

According to the Chapter 2 Article 5 Point 1 (c) of GDPR (General Data Protection Regulation):

1. Personal data shall be:
(c) adequate, relevant and limited to what is necessary in relation the purposes for which they are processed ('data minimisation');

The ‘data minimisation’ principle requires data controllers to minimize the collection of irrelevant information for the analytical target. From table 4-1 , we find 4 European Internet companies collect “friends” information. As far as I am concerned, when the users do not want to share such infortiom, the companies cannot collect. The Interent companies should collect users’ information out of reasonable purpuses, rather than colllect as much as they can.

5.1.3 Discussions of the knowledge graph results — contact us

According to the Chapter 3 Article 13 Point 1 of GDPR (General Data Protection Regulation):

> *1. Where personal data relating to a data subject are collected from the data subject, the controller shall, at the time when personal data are obtained, provide the data subject with all of the following information:*
> *(a) the identity and the contact details of the controller and, where applicable, of the controller’s representative;*

The data controllers shall provide contact details for data subjects. In section 4.5, the “contact us” ways in privacy policy of Chinese and European Internet companies are shown with the knowledge graph and table. Taken together, all sample companies offer at least one way for users to contact them on online privacy issues. Email is the most popular way for both Chinese and European Internet companies because it is continent and easy for companies to manage. The difference is that European Internet companies seems to be more passionate about traditional contact ways such as letter and phone call.

According to the Chapter 3 Article 12 Point 3 & 4 of GDPR (General Data Protection Regulation):

> *3. The controller shall provide information on action taken on a request under Articles 15 to 22 to the data subject without undue delay and in any event within one month of receipt of the request. That period may be extended by two further months where necessary, taking into account the complexity and number of the requests. The controller shall inform the data subject of any such extension within one month of receipt of the request, together with the reasons for the delay. Where the data subject makes the request by electronic form means, the information shall be provided by electronic means where possible, unless otherwise requested by the data subject.*

> *4. If the controller does not take action on the request of the data subject, the controller shall inform the data subject without delay and at the latest within one month of receipt of the request of the reasons for not taking action and on the possibility of lodging a complaint with a supervisory authority and seeking a judicial remedy.*

The data controllers shall take actions for the data subjects' requests in time, no more than one month in any case. Even though the data controllers cannot take actions, they shall inform the data subjects of the possibility of looking for help from the third parties. It is necessary for the Internet companies to provide their contact information. At the same time, it is better to tell users the reply time to ensure the privacy protection effective in a given time, which also helps to improve the users' sense of security. GDPR gives a clear time limit — no more on month, to take actions for the data subjects' requests, so the Internet companies would better set an eligible reply deadline in their privacy policies. In this sense, Chinese Internet companies do better than European Internet companies do.

5.2 Findings of the comparisons and discussions

This section is mainly for summarizing the comparisons and discussion results, in order to provide ideas for following suggestion section. The approach is to list all the comparisons and discussion results, and then summarize them into findings.

In general, whatever in Europe or China, the Internet companies provide privacy policies and put them on the bottom of their official websites. But some Chinese Internet companies do a bad work in this respect. Their privacy policy links are difficult to find and some even do not have privacy policy.

Privacy policies of European Internet companies are greatly affected by GDPR. This can be seen from their privacy policy update time, which is concentrated on May 25th, 2018 - the date of GDPR taken into effect. The privacy policy update time of Chinese Internet companies is separate, and some companies do not show the update time.

For the "collected information" part of privacy policy, the entities extracted from the European Internet companies are more than those of Chinese, and the relationships among entities are more complicated than those of Chinese. Analyzing based on the original corpus, one possible reason is the European Internet companies list the "collected information" in details, some using tables. One principle of personal data processing in GDPR is "Personal data shall be processed lawfully, fairly and in a transparent manner". According to this, the European Internet companies do better in this respect than Chinese Internet companies do.

The “data minimization” principle of GDPR requires data controllers to minimize the collection of personal data. In both Chinese and European Internet companies’ privacy policies, users’ name, email address and other basic personal information are collected. European Internet companies prefer to collect users’ health data, Facebook account and other personal information, compared with Chinese Internet companies. Chinese Internet companies would rather collect personal information such as location and credit card number.

GDPR requires the data controllers shall provide contact details for data subjects. Overall, no matter Chinese or European Internet companies, provide at least one contact way, in which email is the most popular. Compared with Chinese Internet companies, European Internet companies prefer to offer traditional contact methods like postal address and phone number.

GDPR stipulates the data controllers shall take actions for the data subjects’ requests in no more than one month in any case. In the sample, no European Internet company give specific reply time, while half Chinese Internet companies set eligible reply deadline.

To sum up, in the respect of privacy policy, European Internet companies conform better to GDPR than Chinese Internet companies do. Chinese Internet companies have a long way to go. There are merits of Chinese Internet companies, for example, they provide specific reply time in the privacy policy.

5.3 Suggestions based on the findings

The Internet has penetrated into people's lives, and people pay more and more attention to their online privacy protection. Facing users’ privacy protection, enterprises should take the initiative to defend such as setting a sound privacy policy, rather than passively remedy the mistake when an event happen (Jingdong Institute of Law, 2018). Developing a sound privacy policy not only proves the credibility of Internet companies and improves their competitiveness, but also provides guidance for users' privacy protection and eliminate users' privacy concerns, which is the best of both worlds. Based on the above studies, the following suggestions are put forward for the Internet companies:

- Set privacy policy and put it on an obvious location. GDPR requires the data controllers shall provide “transparent information, communication and modalities for the exercise of the rights of the data subject”. Privacy policy is no doubt a perfect way to practice this piece of regulation. Putting the privacy policy at the bottom of the official website is the most common way. Chinese

Internet companies shall adopt the best practice from European Internet companies in this respect.

- Update privacy policy in time. Keeping the privacy policy the same pace with the local laws can avoid the punishment, especially for the multinational enterprises. Showing the update time reflects the enterprises' normalization.
- List the collected personal information in the privacy policy. GDPR stipulates "personal data shall be processed lawfully, fairly and in a transparent manner". So before collecting users' information, let them know. Tables and figures are good ways to show the lists of collected personal information. Chinese Internet companies shall adopt the best practice from European Internet companies in this respect.
- Do not collect personal information irrelevant to the target. "Data minimization" principle of GDPR requires data controllers to minimize the collection of personal data. If the Internet companies do not have reasonable reasons, do not collect information from users. Both Chinese and European Internet companies need to pay attentions on this.
- Provide effective contact methods in the privacy policy and set reply deadline. GDPR requires the data controllers shall provide contact details and take actions for the data subjects' requests in no more than one month in any case. The Internet companies shall be aware that the quality of their solutions to users is related to their competitiveness. European Internet companies must notice this because there is no one of sample company sets reply deadline in the privacy policy.

Chinese Internet companies are not as good as European Internet companies are on many aspect of privacy policy. The main reason is China does not have its own laws or regulations about personal data protection like GDPR. It is time for China to create its own personal data protection laws according to the national conditions.

6 ASSESSMENT OF THIS STUDY

6.1 Theoretical implications and practical contributions

This thesis aims to find the differences between Chinese and European Internet companies' privacy policy, and give suggestions according to GDPR, in order to improve the privacy policy and solve the privacy protection problem in a sense.

About theoretical implications of this study, although there have been comparative studies of privacy policy in different areas such as e-commerce websites and library (Zhou & Wang, 2017; Tian & Xu, 2015), and studies of finding problems of privacy policy (Pollach, 2006), they all stop on a layer of statistics and text explanation. Basically, no study analyzes the privacy policy at an entity level. This thesis is to collect enough privacy policy samples and construct privacy policy knowledge graphs, so that we can recognize the differences from a detailed perspective, which is more specific than the non-entity ones. Additionally, the knowledge graph is also applied and extended in the field of privacy policy. In recent years, constructing knowledge graph of specific domain and improving the technology of constructing knowledge graph are one polular research focus. For example, researchers have constructed health knowledge graph from electronic medical records (Rotmensch et al., 2017). This thesis can be regarded as an application of knowledge graph in the field of privacy policy, and the field of knowledge graph is extended at the same time.

As for practical contributions, the most obvious practical contribution is the same as the aim of this study. As is mentioned above, studying privacy policy can solve the privacy protection problem in a sense. In this way, the users' privacy can be protected in some sense. On the other hand, China has not had personal data protection law like GDPR yet. This study can provide references for the creation of Chinese "GDPR". For example, the editors of Chinese personal data protection laws can set a regulation about "listing the collected personal information in the privacy policy", because we find Chinese Internet companies prefer to generalize the personal data they collected in their privacy policies rather than list and show them in details to users. European Internet companies do better in this respect. The gap has been found, waiting for solving. Overall, privacy policy is important for the users' online privacy protection, and this study hopes to help protecting users' online privacy policy by the comparative study.

6.2 Limitation of this study and future work

It should be pointed out that the entity extraction method of this study is not generalizable to any sample. This study focuses on the comparative study rather than improving the technology, so when we meet the problem that the results of entity extraction is not perfect, we processed them manually and finally got a good result. The simplified process from appendix 6 and 7 to table 4-1 and table 4-2 is by human understanding and picking. This can be worked out when the sample is not too big, but if the entities reach millions or more, manual work can solve nothing.

Based on above limitation, one future work can be put forward: try to improve the entity extraction effect by other technologies. Another prospect is to develop a system to simplify human-reading process of the privacy policy. Through the study, we find the privacy policy is too long for users to read, which has no benefit for users to protect their online privacy. Thus we can develop a system to simplify the reading process of users on the privacy policy, and the privacy policy knowledge graph in this study can be used in the system. Laws and regulations can also be introduced in different areas as standards for users to have an objective judgement. For example, GDPR can be set as the standard for European enterprises' privacy policy.

SUMMARY

The initial motivation of this study is because the writer is disturbed by too many crank calls, and the callers know exactly your personal information such as your family name and your education degree. It is likely that the companies you filled in your personal information online leak your information. Privacy policy is not only a way of company self-discipline, but also a means to protect users' private information. Having a comparative study also helps to find the gap and improve the privacy policy. Knowledge graph can simplify and visualize unstructured data like text data to some extent, and few studies apply knowledge graph in the field of privacy policy. Therefore, “A comparative study of Chinese and European Internet companies’ privacy policy based on knowledge graph” is come up with.

The purpose of this study is to find the similarity and differences between Chinese and European Internet companies, and provide suggestions for the privacy policy of Internet companies with GDPR. To achieve it, the first step, sampling is had with the ranking lists. In the process of looking for privacy policy links, we find some companies do not show their privacy policies, and for some companies, it is hard to find their privacy policies. Moreover, the update time is typically in the beginning or at the end of the privacy policy, which is convenient to count. So we have an overall comparison from these three perspectives: whether there is privacy policy, the location of the privacy policy and the update time of the privacy policy. The results are (1) all sample European Internet companies have privacy policies but some Chinese Internet companies do not have, (2) the privacy policy location of European Internet companies is easier to reach than those of Chinese Internet companies, (3) the update time of European Internet companies’ privacy policy is concentrated on before or after May 25th, 2018, while the update time of Chinese Internet companies is separate.

The following step is the construction of knowledge graphs. Based on literature of knowledge graph, we make the construction process of this study. That is, corpus preprocessing, entity extraction and storing in the graph database, in which the entity extraction is the core and difficult part. Corpus preprocessing is about word segment and part-of-speech tagging, as well as the format edition, all of which can be realized by toolkit “Part-of-speech tagger” and software “UltraEdit’. The process of entity extraction is actually the process of predicting to label the test dataset by the labeled train dataset. What we have to do are labeling the characteristics of train dataset by “BIO” and making characteristic template, as well as editing the format to match the toolkit “CRF++”. The program in “CRF++” can achieve the entity recognition.

The entities extracted from the corpus can be save with the company name as “.csv” file, and imported in the graph database “Neo4j”. The privacy policy knowledge graphs

are generated. We choose two common parts — “collected information” and “contact us” as the corpuses, and construct knowledge graphs separately.

In the discussions, the overall comparison results are directly analyzed with GDPR. Since the privacy policy knowledge graphs of “collected information” contain too many entities, we query in the knowledge graphs and further analyze the results with GDPR. The privacy policy knowledge graphs of “contact us” can also be analyzed directly with GDPR. Through the discussions and analysis, the following findings are concluded: (1) Some Chinese Internet companies do not have privacy policies and it is difficult to find their privacy policy links. (2) European Internet companies’ privacy policies are influenced more by GDPR than Chinese companies. (3) European Internet companies list clearly what personal information they collect in their privacy policies, but Chinese Internet companies like generalizing the personal information they collected. (4) European Internet companies like collecting health data, Facebook account and other personal information, while Chinese Internet companies like collecting personal information such as location and credit card number. (5) Both Chinese and European Internet companies provide at least one contact method. European Internet companies do not give the reply deadline. According to these findings, five suggestions are proposed for the Internet companies on their privacy policies: (1) set privacy policy and put it on an obvious location, (2) update privacy policy in time, (3) list the collected personal information in the privacy policy, (4) do not collect personal information irrelevant to the target, (5) provide effective contact methods in the privacy policy and set reply deadline. From the gap between Chinese and European Internet companies on privacy policies, another constructive suggestion is China shall create its own personal data protection laws according to the national conditions.

In summary, this thesis applies knowledge graph to the privacy policy study, using a comparative perspective to analyze privacy policies in the Internet company industry, and provide promoting suggestions with GDPR. At the same time, this thesis also brings some thinking to the formulation of personal data protection laws in China.

REFERENCES

Alfonseca, E., & Manandhar, S. (2002). An unsupervised method for general named entity recognition and automated concept discovery. In *Proceedings of the 1st international conference on general WordNet, Mysore, India* (pp. 34-43).

Attaran, M., & VanLaar, I. (1999). Privacy and security on the Internet: how to secure your personal information and company data. *Information Management & Computer Security*, *7*(5), 241-247.

Baidu Translation. <https://fanyi.baidu.com/>, retrieved 24.2.2019.

Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., et al. & Martin, M. J. (2005). The universal protein resource (UniProt). *Nucleic acids research*, *33*(suppl_1), D154-D159.

Benassi, P. (1999). TRUSTe: an online privacy seal program. *Communications of the ACM*, *42*(2), 56-57.

Berners-Lee, T. J. (1989). *Information management: A proposal* (No. CERN-DD-89-001-OC).

Berners-Lee, T., Chen, Y., Chilton, L., Connolly, D., Dhanaraj, R., Hollenbach, J., et al. & Sheets, D. (2006). Tabulator: Exploring and analyzing linked data on the semantic web. In *Proceedings of the 3rd international semantic web user interaction workshop* (Vol. 2006, p. 159).

Bikel, D. M., Miller, S., Schwartz, R., & Weischedel, R. (1998). Nymble: a high-performance learning name-finder. *arXiv preprint cmp-lg/9803003*.

Bizer, C. (2003). D2r graph-a database to rdf graphing language. < https://www.researchgate.net/publication/2914671_D2R_MAP_--_A_Database_to_RDF_Mapping_Language >, retrived 7.6.2019.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., & Taylor, J. (2008). Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data* (pp. 1247-1250).

Brodie, C. A., Karat, C. M., & Karat, J. (2006). An empirical study of natural language parsing of privacy policy rules using the SPARCLE policy workbench. In *Proceedings of the second symposium on Usable privacy and security* (pp. 8-19).

Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka, E. R., & Mitchell, T. M. (2010). Toward an architecture for never-ending language learning. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*.

Chao, J. I., Perevedentseva, E., Chung, P. H., Liu, K. K., Cheng, C. Y., Chang, C. C., & Cheng, C. L. (2007). Nanometer-sized diamond particle as a probe for biolabeling. *Biophysical journal*, *93*(6), 2199-2208.

Chen, Shaojian, (2017) *Neo4j full-stack development*. Publishing House of Electronics Industry, Beijing.

Chinchor, N., & Robinson, P. (1997). MUC-7 named entity task definition. In *Proceedings of the 7th Conference on Message Understanding* (Vol. 29, pp. 1-21).

Collier, D. (1993). The comparative method. *Political Science: The State of Discipline II, Ada W. Finifter, ed., American Political Science Association*.

Collins, M., & Singer, Y. (1999). Unsupervised models for named entity classification. In *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

Cook, D. L., & Coupey, E. (1998). Consumer behavior and unresolved regulatory issues in electronic marketing. *Journal of business research*, *41*(3), 231-238.

Craven, M., & Kumlien, J. (1999). Constructing biological knowledge bases by extracting information from text sources. In *ISMB* (Vol. 1999, pp. 77-86).

CRF++: Yet Another CRF toolkit. <https://taku910.github.io/crfpp/>, retrieved 10.3.2019.

Cucchiarelli, A., & Velardi, P. (2001). Unsupervised named entity recognition using syntactic and semantic contextual evidence. *Computational Linguistics*, *27*(1), 123-131.

Cucerzan, S. (2007). Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

Dalton, J., Dietz, L., & Allan, J. (2014). Entity query feature expansion using knowledge base links. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval* (pp. 365-374).

Davies, J., Fensel, D., & Van Harmelen, F. (Eds.). (2003). *Towards the semantic web: ontology-driven knowledge management*. John Wiley & Sons.

DeRose, P., Shen, W., Chen, F., Lee, Y., Burdick, D., Doan, A., & Ramakrishnan, R. (2007). DBLife: A community information management platform for the database research community. In *CIDR* (pp. 169-172).

Deshpande, O., Lamba, D. S., Tourn, M., Das, S., Subramaniam, S., Rajaraman, A., et al. & Doan, A. (2013). Building, maintaining, and using knowledge bases: a report from the trenches. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data* (pp. 1209-1220).

Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A. M., Shaked, T., et al. & Yates, A. (2004). Web-scale information extraction in knowitall:(preliminary results). In *Proceedings of the 13th international conference on World Wide Web* (pp. 100-110).

Etzioni, O., Cafarella, M., Downey, D., Popescu, A. M., Shaked, T., Soderland, S., et al. & Yates, A. (2005). Unsupervised named-entity extraction from the web: An experimental study. *Artificial intelligence*, *165*(1), 91-134.

EU GDPR.ORG. Timeline of Events. <https://eugdpr.org/the-process/timeline-of-events/>, retrieved 3.3.2019.

Flavián, C., & Guinalíu, M. (2006). Consumer trust, perceived security and privacy policy: three basic elements of loyalty to a web site. *Industrial Management & Data Systems*, *106*(5), 601-620.

Gambell, T., & Yang, C. (2006). Word segmentation: Quick but not dirty. *Unpublished manuscript*.

General Data Protection Regulation. <http://data.consilium.europa.eu/doc/document/ST-5419-2016-REV-1/en/pdf>, retrieved 13.3.2019.

GeoNames. <https://www.geonames.org/>, retrieved 6.3.2019.

Getting Started In: Sequence Labeling. <https://nlpers.blogspot.com/2006/11/getting-started-in-sequence-labeling.html>, retrieved 11.3.2019.

Gindin, S. E. (1999). Creating an Online Privacy Policy. *Preventive L. Rep.*, *18*, 10.

Gole, S. (2015). Part-of-speech tagging using OpenNLP. <https://blog.thedigitalgroup.com/part-of-speech-tagging-using-opennlp>, retrieved 20.3.2019.

Gregorowicz, A., & Kramer, M. A. (2006). Mining a large-scale term-concept network from wikipedia. < https://www.researchgate.net/publication/200773446_Mining_a_Large-Scale_Term-Concept_Network_from_Wikipedia >, retrived 7.6.2019.

Hakkani-Tür, D., Celikyilmaz, A., Heck, L., Tur, G., & Zweig, G. (2014). Probabilistic enrichment of knowledge graph entities for relation detection in conversational understanding. In *Fifteenth Annual Conference of the International Speech Communication Association*.

Hassanzadeh, O., & Consens, M. P. (2009). Linked Movie Data Base. In *LDOW*.

Hill Rebecca. (2018). Max Schrems is back: Facebook, Google hit with GDPR complaint. The Register. <https://www.theregister.co.uk/2018/05/25/schrems_is_back_facebook_google_get_served_gdpr_complaint/?utm_source=tuicool&utm_medium=referral>, retrieved 7.10.2018.

Hixon, B., Clark, P., & Hajishirzi, H. (2015). Learning knowledge graphs for question answering through conversational dialog. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*(pp. 851-861).

Informilo. Europe's 25 Hottest Internet Companies. <http://www.informilo.com/2010/12/europe-s-25-hottest-internet-companies/>, retrieved 24.2.2019.

Jingdong Institute of Law. (2018). *Comments on EU General Data Protection Regulations (GDPR) and Practical Guideline*. Law Press, Beijing, China.

Kang, J. (1997). Information privacy in cyberspace transactions. *Stan. L. Rev.*, *50*, 1193.

Karjoth, G., & Schunter, M. (2002). A privacy policy model for enterprises. In *Proceedings 15th IEEE Computer Security Foundations Workshop. CSFW-15* (pp. 271-281). IEEE.

Koprowski, G. (1995). The Electronic Watchdog'. *American Demographics (July/August)*, *4*(9).

Kwon, O. (2010). A pervasive P3P-based negotiation mechanism for privacy-aware pervasive e-commerce. *Decision Support Systems*, *50*(1), 213-221.

Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Lappalainen, J. (2017). Consent as a legal ground for the processing of personal data: changes and challenges in the context of the EU general data protection reform.

Larose, D. T., & Larose, C. D. (2014). *Discovering knowledge in data: an introduction to data mining*. John Wiley & Sons.

Leacock, C., Chodorow, M., Gamon, M., & Tetreault, J. (2010). Automated grammatical error detection for language learners. *Synthesis lectures on human language technologies*, *3*(1), 1-134.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., et al. & Bizer, C. (2015). DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, *6*(2), 167-195.

Li, H., Sarathy, R., & Xu, H. (2011). The role of affect and cognition on online consumers' decision to disclose personal information to unfamiliar online vendors. *Decision Support Systems*, *51*(3), 434-445.

Li, J., & Hou, L. (2017). A review of knowledge graph research. *Journal of Shanxi University*.

Li, K. (2018). *Construction and research of knowledge graph of obstetrics*. Master's thesis. Zhengzhou University, Zhengzhou.

Liu, B., Wan, L., & Li, Y. (2016). A review of privacy protection based on privacy policy in network environment. *Information studies: Theory & Application*, 39(9), 134r139.

Liu, C. (2018). *Research on medical knowledge search based on knowledge graph*. Master's thesis. Zhejiang University of Technology, Zhejiang.

Liu, F., Wilson, S., Story, P., Zimmeck, S., & Sadeh, N. (2018). Towards Automatic Classification of Privacy Policy Text.

Marchiori, M., Cranor, L., Langheinrich, M., Presler-Marshall, M., & Reagle, J. (2002). The platform for privacy preferences 1.0 (P3P1. 0) specification. *World Wide Web Consortium Recommendation REC-P3P-20020416*.

Ma, Y., Crook, P. A., Sarikaya, R., & Fosler-Lussier, E. (2015). Knowledge graph inference for spoken dialog systems. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5346-5350). IEEE.

McBride, B. (2004). The resource description framework (RDF) and its vocabulary description language RDFS. In *Handbook on ontologies* (pp. 51-65). Springer, Berlin, Heidelberg.

McCallum, A., & Li, W. (2003). Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4* (pp. 188-191). Association for Computational Linguistics.

McCallum, A., Freitag, D., & Pereira, F. C. (2000). Maximum Entropy Markov Models for Information Extraction and Segmentation. In *Icml* (Vol. 17, No. 2000, pp. 591-598).

Michelfelder, D. P. (2001). The moral value of informational privacy in cyberspace. *Ethics and Information Technology*, *3*(2), 129-135.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39-41.

Moens, M., & Steedman, M. (1987). Temporal ontology in natural language. In *Proceedings of the 25th annual meeting on Association for Computational Linguistics* (pp. 1-7). Association for Computational Linguistics.

Mun, M., Hao, S., Mishra, N., Shilton, K., Burke, J., Estrin, D., et al. & Govindan, R. (2010). Personal data vaults: a locus of control for personal data streams. In *Proceedings of the 6th International COnference* (p. 17).

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, *30*(1), 3-26.

NPC China. (2018). The 13th Legislative planning of the Standing Committee of the National People's Congress. <http://www.npc.gov.cn/npc/xinwen/2018-09/10/content_2061041.htm>, retrieved 7.10.2018.

Ohlhorst, F. (2012). *Big data analytics: turning big data into big money* (Vol. 65). John Wiley & Sons.

Ontotext. What is knowledge graph?. <https://www.ontotext.com/knowledgehub/fundamentals/what-is-a-knowledge-graph/>, retrieved 3.3.2019.

Organisation for Economic Co-operation and Development. (2002). *OECD guidelines on the protection of privacy and transborder flows of personal data*. OECD Publishing.

Palmer, M. (2019). Goodbye Blippar, welcome back Layar?. Sifted. <https://sifted.eu/articles/goodbye-blippar-welcome-back-layar/>, retrieved 9.3.2019.

Paulheim, H. (2017). Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic web*, *8*(3), 489-508.

Pollach, I. (2005). A typology of communicative strategies in online privacy policies: Ethics, power and informed consent. *Journal of Business Ethics*, *62*(3), 221.

Pollach, I. (2006). Privacy statements as a means of uncertainty reduction in WWW interactions. *Journal of Organizational and End User Computing (JOEUC)*, *18*(1), 23-49.

Pollach, I. (2007). What's wrong with online privacy policies?. *Communications of the ACM*, *50*(9), 103-108.

Pujara, J., Miao, H., Getoor, L., & Cohen, W. (2013). Knowledge graph identification. In *International Semantic Web Conference* (pp. 542-557). Springer, Berlin, Heidelberg.

Qi, G., Gao, H., & Wu, T. (2017). Research progress of knowledge graph. *Technology Intelligence Engineering*, *3*(01), 4-15.

Qiao, L., Yang, L., Hong, D., Yao, L., & Zhiguang, Q. (2016). Knowledge graph construction techniques. *Journal of Computer Research and Development*, *53*(3), 582-600.

Robinson, I., Webber, J., & Eifrem, E., (2013) *Graph Databases*. O'Reilly Press, Sebastopol.

Rotmensch, M., Halpern, Y., Tlimat, A., Horng, S., & Sontag, D. (2017). Learning a Health Knowledge Graph from Electronic Medical Records. Scientific Reports, 7(1). doi:10.1038/s41598-017-05778-z.

Savage, S. J., & Waldman, D. M. (2015). Privacy tradeoffs in smartphone applications. *Economics Letters*, *137*, 171-175.

Shao, Y. (2017). *Construction and application of knowledge g for industrial products based on Web*. Master's thesis. Shenyang Aerospace University, Shenyang.

Shen, Q. (2017). Research on Privacy Protection Policy of Websites in China: Content Analysis Based on 49 Websites. *Journalism Quarterly*. *39*(12),107-114.

Shinzato, K., & Torisawa, K. (2004). Acquiring hyponymy relations from web documents. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*.

Sigurbjörnsson, B., & Van Zwol, R. (2008). Flickr tag recommendation based on collective knowledge. In *Proceedings of the 17th international conference on World Wide Web* (pp. 327-336).

Singhal, A. (2012). Introducing the knowledge graph: things, not strings. *Official google blog*, *5*.

Steiner, T., Verborgh, R., Troncy, R., Gabarro, J., & Van de Walle, R. (2012). Adding realtime coverage to the google knowledge graph. In *11th International Semantic Web Conference (ISWC 2012)*.

Story, P., Zimmeck, S., Ravichander, A., Smullen, D., Wang, Z., Reidenberg, J., et al. & Sadeh, N. (2019). Natural Language Processing for Mobile App Privacy Compliance.

Su, Y., Yang, S., Sun, H., Srivatsa, M., Kase, S., Vanni, M., & Yan, X. (2015). Exploiting relevance feedback in knowledge graph search. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135-1144).

Suchanek, F. M., Kasneci, G., & Weikum, G. (2007). Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web* (pp. 697-706).

Tang, Y., & Lai, X. (2018). A research on social media privacy policy text: a comparative analysis based on Facebook and WeChat. *News and Writing*. *2018*(08), 31-37.

Tankard, C. (2016). What the GDPR means for businesses. *Network Security*, *2016*(6), 5-8.

Tennakoon, C., Zaki, N., Arnaout, H., Elbassuoni, S., El-Hajj, W., & Al Jaberi, A. (2019). Biological Knowledge Graph Construction, Search, and Navigation. In *Leveraging Biomedical and Healthcare Data* (pp. 107-120). Academic Press.

Tesfay, W. B., Hofmann, P., Nakamura, T., Kiyomoto, S., & Serna, J. (2018). PrivacyGuide: towards an implementation of the EU GDPR on internet

privacy policy evaluation. In *Proceedings of the Fourth ACM International Workshop on Security and Privacy Analytics* (pp. 15-21).

TheFreeDictionary. Dot com company. <https://www.thefreedictionary.com/dot+com+company>, retrieved 2.3.2019.

Tian, S., & Xu, C. (2015). Text analysis and enlightenment of foreign library user privacy protection guidelines. *Library and Information Service*. *116*(18), 61-66.

Tsuruoka, Y. (2005). A part-of-speech tagger for English. <http://www.nactem.ac.uk/tsuruoka/postagger/>, retrieved 10.3.2019.

Tsuruoka, Y., & Tsujii, J. I. (2005). Bidirectional inference with the easiest-first strategy for tagging sequence data. In *Proceedings of the conference on human language technology and empirical methods in natural language processing* (pp. 467-474). Association for Computational Linguistics.

Voigt, P., & Von dem Bussche, A. (2017). The EU General Data Protection Regulation (GDPR). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*.

Wang, C., Ma, X., Chen, J., & Chen, J. (2018). Information extraction and knowledge graph construction from geoscience literature. *Computers & Geosciences*, *112*, 112-120.

Wang, Z., Zhang, J., Feng, J., & Chen, Z. (2014). Knowledge graph and text jointly embedding. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1591-1601).

Westin, A. F. (1967). Privacy and freedom Atheneum. *New York*, *7*, 431-453.

Wu, W., Li, H., Wang, H., & Zhu, K. Q. (2012). Probase: A probabilistic taxonomy for text understanding. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data* (pp. 481-492).

XinhuaNet. (2018). List of the top 100 Internet companies in China for 2018. <http://www.xinhuanet.com/fortune/2018-07/27/c_129921892.htm?baike>, retrieved 24.2.2019.

Xu, H., Dinev, T., Smith, J., & Hart, P. (2011). Information privacy concerns: Linking individual perceptions with institutional privacy assurances. *Journal of the Association for Information Systems*, *12*(12), 798.

Yahya, M., Barbosa, D., Berberich, K., Wang, Q., & Weikum, G. (2016). Relationship queries on extended knowledge graphs. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining* (pp. 605-614).

Zhang, Y. (2017). *The Comparative study of Chinese and foreign big data enterprises' privacy policy*. Master's thesis. Shanxi University, Shanxi.

Zhou, S., & Wang, W. (2017). A comparative study on privacy policies of e-commerce websites in china and America: a case study of Alibaba and Amazon. *Modern Information*, *37*(1), 137-141.

APPENDICES

Appendix 1 The privacy policy corpus of Chinese Internet companies

	Company name	Privacy policy link
1	Alibaba Group	https://rule.1688.com/policy/privacy.html
2	Tencent	https://privacy.qq.com/
3	Baidu	http://privacy.baidu.com/detail?id=288
4	JD.com	https://about.jd.com/privacy/
5	NetEase	http://gb.corp.163.com/gb/legal.html
6	Sina	http://corp.sina.com.cn/chn/sina_priv.html
7	Sohu	http://corp.sohu.com/s2007/privacy/
8	Meutuan	https://portal-portm.meituan.com/webpc/protocolmanage/privacy
9	360 Total Security	http://www.360.cn/privacy/v3/360daohang.html
10	Mi	https://www.mi.com/about/privacy/
11	Bytedance	https://www.bytedance.com/policy/#privacy
12	Wangsu	https://www.wangsu.com/law/privacy.html
13	58 Group	https://about.58.com/395.html?utm_source=sem-360-pc&spm=14572675192.3609375071
14	Kingsoft	http://www.wps.cn/privacy/privacyprotect/
15	Trip.com	https://pages.trip.com/service-guideline/privacy-policy-en-us.html
16	Meitu	https://www.meitu.com/services/privacy.html
17	Suning	https://help.suning.com/page/id-281.htm
18	Autohome	https://www.autohome.com.cn/about/falv.html?f=index&pvareaid=21256822772
19	Migu	https://passport.migu.cn/portal/privacy/protocol
20	37 Interactive Entertainment	http://www.37wan.net/html/privacy.html

Appendix 2 The privacy policy corpus of European Internet companies

	Company name	Privacy policy link
1	Gameforge	https://agbserver.gameforge.com/enGB-Privacy-GF-Portal.html
2	Layar	https://www.layar.com/legal/privacy-policy/
3	Shazam	https://www.apple.com/legal/privacy/en-ww/
4	Spotify	https://www.spotify.com/uk/legal/privacy-policy/
5	Unity	https://unity3d.com/legal/privacy-policy
6	Wonga	https://www.wonga.com/privacy-policy
7	Getjar	https://www.getjar.com/info/privacy/
8	Wooga	https://www.wooga.com/privacy-policy/
9	SoundCloud	https://soundcloud.com/pages/privacy
10	Befunky	https://www.befunky.com/privacy/
11	Vente-privee	https://secure.uk.vente-privee.com/registration/PrivacyPolicy?CountryCode=EN#pp_cookies
12	Yandex	https://yandex.com/legal/privacy/
13	Tradeshift	https://tradeshift.com/privacy-policy
14	Rovio Mobile	http://www.rovio.com/privacy
15	Fotolia	https://www.adobe.com/privacy.html
16	Ooyala	https://www.ooyala.com/privacy
17	Klarna	https://www.klarna.com/uk/privacy-policy/
18	Badoo	https://badoo.com/en/privacy/
19	AVG	https://www.avg.com/en-gb/privacy
20	Criteo	https://www.criteo.com/privacy/

Appendix 3 The click times to reach the privacy policy of Chinese Internet companies

	Company name	The steps to reach the privacy policy	The click times
1	Alibaba Group	Home>Privacy Policy	2
2	Tencent	Privacy Protection Platform	1
3	Baidu	Privacy Protection Platform	1
4	JD.com	Home>Privacy Policy	2
5	NetEase	Home>Privacy Policy	2
6	Sina	Home>Privacy Protection	2
7	Sohu	Home>Corporation Introduction>Sohu Corporation>Protect Privacy	4
8	Meutuan	Home>Privacy Policy	2
9	360 Total Security	Privacy Protection Platform	1
10	Mi	Home>Privacy Policy	2
11	Bytedance	Home>Privacy Policy	2
12	Wangsu	Home>Privacy Clause	2
13	58 Group	Home>Privacy Right Clause	2
14	Kingsoft	Home>Service Clause>Privacy Protection Policy	3
15	Trip.com	Home>Privacy Statement	2
16	Meitu	Home>Privacy Policy	2
17	Suning	Home>Privacy Policy	2
18	Autohome	Home>Privacy Policy	2
19	Migu	Home>Privacy Right Policy	2
20	37 Interactive Entertainment	Home>Privacy Policy	2

Appendix 4 The click times to reach the privacy policy of European Internet companies

	Company name	The steps to reach the privacy policy	The click times
1	Gameforge	Home>Privacy	2
2	Layar	Home>Legal>Privacy Policy	3
3	Shazam	Home>Privacy> Privacy Policy	3
4	Spotify	Home>Privacy Policy	2
5	Unity	Home>Privacy Policy	2
6	Wonga	Home>Privacy	2
7	Getjar	Home>Privacy	2
8	Wooga	Home>Privacy Policy	2
9	SoundCloud	Home>Privacy	2
10	Befunky	Home>Privacy Policy	2
11	Vente-privee	Home>Privacy and Cookies Policy	2
12	Yandex	Home>Privacy Policy	2
13	Tradeshift	Home>Privacy Policy	2
14	Rovio Mobile	Home>Privacy Notice	2
15	Fotolia	Home>Privacy	2
16	Ooyala	Home>Privacy	2
17	Klarna	Home>Privacy Statement>Privacy Policy	3
18	Badoo	Home> Privacy	2
19	AVG	Home>Privacy	2
20	Criteo	Home>Privacy Policy	2

Appendix 5 The allocation of train dataset and test dataset

	Chinese Internet companies' privacy policy		**European** Internet companies' privacy policy	
	Test dataset	**Train dataset**	**Test dataset**	**Train dataset**
1	Alibaba Group	NetEase	Gameforge	Layar
2	Tencent	Sohu	Spotify	Unity
3	Baidu	Meutuan	Befunky	SoundCloud
4	JD.com	58 Group	Shazam	Vente-privee
5	Sina	Kingsoft	Wonga	Getjar
6	360 Total Security	Meitu	Rovio Mobile	Tradeshift
7	Mi	Suning	Badoo	Fotolia
8	Bytedance	Autohome	Criteo	Ooyala
9	Wangsu	Migu	Wooga	Klarna
10	Trip.com	37 Interactive Entertainment	Yandex	AVG

Appendix 6 The entities shared by European companies

The number of companies (x)	**Entities shared by x companies**
8	information
7	IP address
6	time
5	your name, your device
4	friends, gender
3	advertising, system, personal data, device, date, your consent, name, website, email address, personal information, address, cookies
2	messages, language, browser type, purchase, marketing communications, Service, your email address, country, order, option, Facebook, image, advertisements, our website, interest, technologies, files, phone number, software, device identifiers, forums, number, special categories, your health, account, contact details, application, technical data, example, political opinions, contacts, form, Services, version, accordance, user account

Appendix 7 The entities shared by Chinese companies

The number of companies (x)	Entities shared by x companies
8	personal information
6	your name, email address
5	name, other information
4	use
3	IP address, information, gender, collection, type
2	clickstream data, operating system you use, Internet browser, someone, postal address, telephone number, identifiable information, updates, your account, device model, log, operating system version, services, browser type, personal data, email communications, contact details, Information, mobile phone number, IP, ID card, IP addresses, Log information, Location information, delivery address, device, order, birthday, credit card number, cookies.