哈爾濱工業大學 *Linköping University*

Thesis Proposal

A Behavior-based Optimization of a Search Query for Patents

Postgraduate	**Sijin Cheng**
Student Number	**16S137031**
Workplace of Internship	**IamIP**
HIT Supervisor	**Prof. Tianyi Zang**
Partner Univ. Supervisor	**Prof. Marco Kuhlmann**
Industry Supervisor	**Mr. Falah Hosini**
Time of Submitted	**2018/3/21**

Contents

1 The purpose and significance of thesis topic 1
1.1 Project title 1
1.2 The subject source 1
1.3 State of the art 1
1.3.1 Research status of text mining 1
1.3.2 The development of patent mining 2
1.4 Main content of this topic 3
2 Research plan 4
2.1 Requirement analysis 4
2.1.1 The functional requirements 4
2.1.2 The un-functional requirements 5
2.2 Technical method 6
2.3 Implementation conditions 7
2.3.1 Technical condition 7
2.3.2 Experiment condition 7
2.4 Exist problems and technical keys 8
2.4.1 Exist problems 8
2.4.2 Technical keys 8
2.5 Expected goal 12
3 Schedule and work plan 13
3.1 Schedule 13
3.2 Work plan and implementation 14
Major references 15
Evaluations of supervisors 16

1 The purpose and significance of thesis topic

1.1 Project title

A Behavior-based Optimization of a Search Query for Patents

1.2 The subject source

This is a project that is supplied by a patent management company, IamIP. To catch the important trends in technology or materials, the company need to keep a continuous watch on the patent activities within their field. This is normally done via a search alert that signals the user when any new patents relevant to the search query are published. IamIP's Platform is an innovative tool for patent search and patent management. There are more than 100 million patents from around the world on the IamIP's database, and the data update every week by adding up to 100,000 new patent. The platform provides excellent patent search possibilities. By setting IamIP Search Alert, the customer can define specific searches to capture the most up-to-date, relevant competitive Intelligence for the business.

However, there are several factors influence the effect of the search alert now. The first is that choosing the right words to search with is difficult so that the search query cannot accurately locate the user's intent[1], the reason of it is that the structure of patent documents is complex and the use of a specific terminology is not always commonly known or used. The second, technologies change, new competitors enter the market, new materials appear, and consumer behavior might change, too, which may be difficult to be observed by the set Search Alert. It is therefore desirable to continuously update and optimize the search query for a search alert, to adapt to observed changes in the response of the search alert owner to newly captured patents.

1.3 State of the art

1.3.1 Research status of text mining

Text data is a kind of unstructured data, one of the simplest forms of data that

humans use. Unstructured text is easily understood and perceived by humans, but it is much more difficult for machines to process. Text mining is the task of extracting unknown, understandable, and potentially meaningful information from text, it covers information retrieval, natural language processing, data mining, machine learning and many other applications[2]. For many text mining tasks, keyword extraction is crucial. Because only the keywords can represent the corresponding documents well, the task of text mining can achieve higher quality[3]. Keyword extraction has been widely studied by many researchers.

Li Juanzi et al.[4] proposed a multi-strategy keyword extraction method based on TF/IDF. This method selects a set of candidate keywords and then defines the features based on their morphological characteristics and contextual information. In addition, some strategies are proposed to correct the incomplete words obtained from participle, and the unknown potential keywords are found in the news literature. The experimental results show that the proposed method has better performance than the baseline method. Rui Wang et al.[5] proposed to embed the word vector as an external repository for keyword extraction and generation method, this method not only considers the extracted from the text keywords, also considered again does not appear in the text of the abstract class. The evaluation results show that the use of word embedding vectors is a simple and effective method for the extraction and generation of keywords. Rada Mihalcea and Paul Tarau[6] introduced TextRank, a graphics-based text processing sorting model, they showed and evaluated how this model can be used for keyword extraction. The results show that even without language knowledge, the accuracy of TextRank can be compared with the most advanced algorithms proposed previously.

1.3.2 The development of patent mining

Patent mining can be divided into patent metadata mining and patent text mining. The former is more mature in methodology and analytical techniques. However, much important information is hidden in the patent description text[3]. Recent years, many researchers conducted patent mining research.

Jie Hu proposed a patent keyword extraction algorithm (PKEA) based on the distribute Skip-gram, and they provided an indirect way to evaluate the effectiveness of extracting meaningful keywords by using information gain. S. Don and Dugki Min[7] proposed a system for the automatic categorization of the patent, they make feature selection based on the term frequency and simplify the feature using information gain (IG). Then random forest (RF), SVM (SVM) and naive Bayes (NB) classifier were used. It is found that the semantic structure information

of patent documents is an important feature of the classification document. Junegak Joung et al.[8] proposed a patent analysis technique based on the keyword model to monitor emerging technologies. After identifying keywords by using text mining tools and technology, they built a keyword context matrix and then identified the relationship between the keywords in the matrix transformation. The patent document can be clustered by adopting hierarchical clustering algorithm. The results show that emerging technologies can be monitored by identifying clusters of technical keywords. Chao-Chan Wu[9] put forward a kind of weighted based on the study of key patent network (WKPN) method, the process of generating WKPN including patent keyword extraction, the calculation of the weighted value of each keyword, the establishment of the similarity matrix and accurate network building. In addition, quantitative indexes are also recommended to analyze the technical influence of WKPN. The results show that this method can not only improve the efficiency and effectiveness of patent analysis but also recognize the development trend and development trend of the emerging technology field.

To our knowledge, no previous work has studied the application of patent mining information to optimize search results. This paper will capture the user's preferences by exploring the characteristics of the patent, and optimize give more relevant search result by generating new search string.

1.4 Main content of this topic

In this paper, we are trying to optimize the existing search query for the search alert by giving interactively query suggestions by learning the user behavior model and present a ranked list of the patents searched using the new query. The contents of this paper are summarized as follows:

(1) Model the user behavior. Firstly, we need to collect the user’s feedback on the query results and set up with user behavior scoring rules. Then we can abstract the features and analyze the information to form a valuable user model.

(2) Generate search suggestion. We need to establish the relationship between user model and patent information based on the suggestion strategy. At last, the query suggestion is given based on the user preference.

(3) Rank the query results. After we query the results based on the new query, we could calculate the similarity of the search string and the new query results.

(4) Evaluation. Evaluating whether the introduction of user preference can reduce the proportion of the irrelevant search results.

2 Research plan

2.1 Requirement analysis

2.1.1 The functional requirements

1. Design the algorithm to analyze the user's preference based on the feedback data.
2. Give the query suggestions for the existing search alert
3. Give a rank list of results based on the new query
4. Evaluate the usefulness of the suggested changes

System Model

At first, we draw a general system model diagram according to the functional requirements, as shown in Figure 2.1.

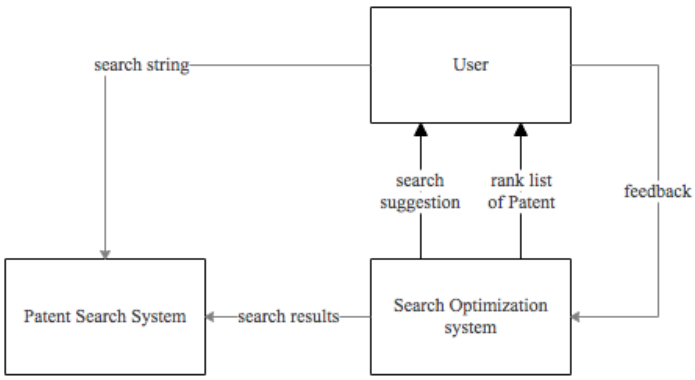


Figure 2.1 System model diagram

Use case diagram

In this project, the patent search system is the existing system in the company IamIP, we could access and reuse the function of it, so we can spend more energy on designing algorithm and focus on search optimization subsystem. The use case diagram for search optimization system is shown in Figure 2.2.

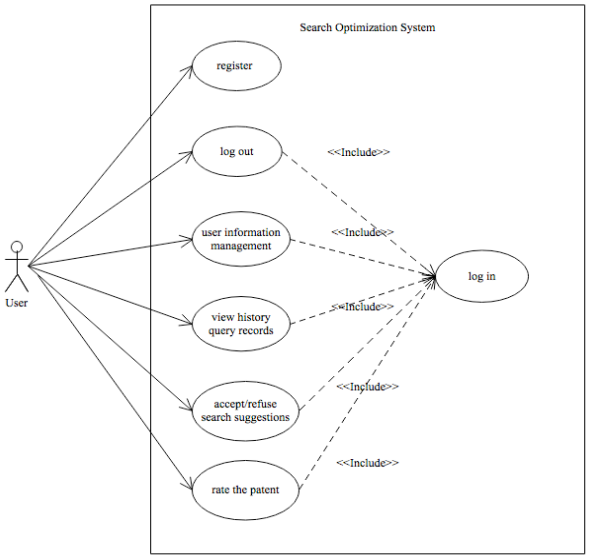


Figure 2.2 Use case diagram

System function structure

After analyzing the use cases, we can divide the function of the search optimization subsystem into three sub-modules: Data preprocess module, Search suggestion module, Ranking, and evaluation module. The detailed functions in each module are as follows.

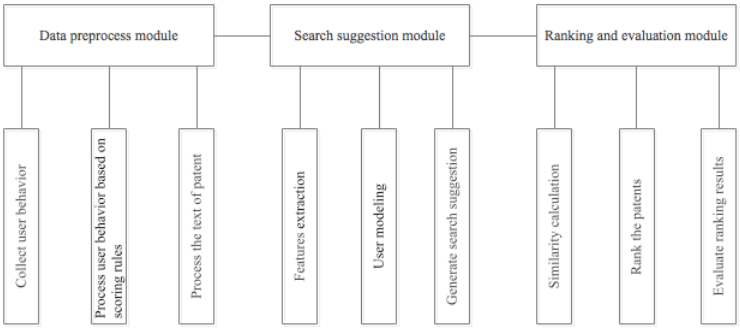


Figure 2.3 Function structure diagram

2.1.2 The un-functional requirements

For this system, non-functional requirements mainly include high efficiency, safety, ease of use and integration.

(1) Easy-to-use: the system needs to be easily operated, which is easy to understand and easy to learn.

(2) Efficiency: the customer requires quick response and minimizes the blocking and errors caused by the system.

(3) Security: it is necessary to protect the privacy of users when the system is

established.

(4) Maintainable: the system should be integrated with other systems easily and the system administrators can maintain it takes little effort.

2.2 Technical method

The process of this project includes four layers:

Data collection

To introduce the user interest model, we need first collect the user's behavior. In this project, user behavior includes scoring the search results according to the degree of interest and throwing the uninterested patents in the trash box. In order to facilitate processing, we set up with user behavior ranking rules[1]. It will inevitably have some subjectivity, but this rule conforms to the behavior characteristics of the majority of users, also can rapidly advance projects.

Table 2-1 Ranking rule

Rank	Patent
1	Irrelevant patent
2	Low related patent
3	High related patent

Data preprocessing

The information of a complete patent information includes: {Patent Title, Abstract, Publication(s), Applicant(s), International filing date, Priority(s), Inventor(s), IPC, CPC, Simple family member(s), Claims, Ranking, Drawing}.

In this project, we only focus on Title and abstracts of the Patent. The title and abstract of the patent is natural language, which the computer can't understand. Therefore, the first problem is to properly represent the text in the computer. For this purpose, it should contain enough information to reflect the features of the text without being too complicated for the learning algorithm. We need a series of pretreatments for the particularity of patent documents.

Search suggestion

(1) User model

This paper only uses user's feedback as the main source of user preference, without considering the user's other attributes. We can abstract the features to form a valuable user model.

(2) Suggestion

We need to establish the relationship between user model and patent information based on the suggestion strategy. At last, the query suggestion is given based on the user preference.

(3) Ranking list

Calculate the similarity of the candidate patents and the new search string, rank the candidate patents based on the similarity

Evaluation

Compared the new results with the original query results (check whether the proportion of irrelevant results is decreased), compare the ranking with the real score (available from the customer).

2.3 Implementation conditions

2.3.1 Technical condition

The technical requirements for the development of this project are as follows:

1. Text representation and feature extraction
2. Similarity calculation
3. Algorithm design
4. Evaluation metrics

2.3.2 Experiment condition

Python 3.6

There are several reasons that we choose Python as the development language:

1. Easy to use: Python is a simple and elegant language with complete documentation. When writing a program in Python, there is no need to consider the underlying details such as how to use memory
2. Powerful libraries: The Python standard library is very large and it can help with a variety of tasks, including regular expressions, natural language processing, document generation, unit testing, databases and more.
3. Free and open source: Python is one of the Free/ Libre and Open Source Software (FLOSS), users are free to read its source code or make changes to it
4. Portability: Due to it is an open source software, Python has been ported on many platforms
5. Integration: The existing search platform in IamIP is also based on Python, which makes it easier for the integration of other in the company.

Scikit-learn 0.19.1

Scikit-learn is an open source, simple and efficient tool for data mining and data analysis, which is built on NumPy, SciPy, and matplotlib. It is accessible to

everybody and easy to learn because it contains complete documentation and many machine learning algorithms.

2.4 Exist problems and technical keys

2.4.1 Exist problems

1. The dataset is small

When calculating the feature weights, due to the evaluation function is based on the statistics, one of the main flaws is that it need to use a very large training set to get almost all of the features for classification is crucial. It takes a lot of time and space resources, and building such a huge training set is a daunting task. In real applications, however, considering the working efficiency, we will not also do not have enough resources to build such a large training set, the result is: the selected weights and even higher characteristic, may be little use for classification, it will interfere with the correct classification; However, the really useful features are given lower weight because of the low frequency of occurrence, and even deleted when the feature space dimension is reduced.

2. Lack of unified evaluation criteria

The traditional information retrieval system has perfect performance evaluation standard. However, for this system, we can't simply apply the evaluation standard of the traditional information retrieval system. For this system, the choice of evaluation index is the key. The performance of the system should be evaluated from the evaluation indexes such as high efficiency, safety, ease of use and maintainability. However, how to measure the quality of the evaluation of service has not found an objective, fair and uniform quantitative evaluation standard. In the real world, users seldom actively participate in information feedback. As a result, a command push by developing a prototype system, and employ volunteer who use the system to collect user feedback, but this way for the validation of the average user data sets is too small, time-consuming, and based on the feedback of performance evaluation is subjective.

2.4.2 Technical keys

1. Text presentation

Since the text is unstructured data, it is necessary to the original text into a structured information that computer can identify and process.

Tokenization

Word is the smallest meaningful constituent unit in the natural language. Tokenization refers to dividing sentences into single words, and word segmentation results can be used for further processing, so the correct automatic word segmentation is the basis of the correct Natural Language Processing.

Filtering

Filtering refers to deleted some words from the document, stop words and punctuation are usually to be filtered. Stop words are those occur frequently in the text, but without actual meaning words such as determiners ("the", "a", "an", "that", "those", etc.) and prepositions ("on", "at", "above"). Similarly, those words that appear frequently in texts do not explain information to distinguish different documents also can be deleted from documents. [10]

Stemming

Stemming is to consider the morphological changes of words, namely the grouping of the inflected forms of a word can be regarded as a word.

Vector Space Model

We could represent text information by quantifying the special words extracted from the text and establish its mathematical model. The computer can realize the recognition of text through the calculation and operation of this model. Vector space models are commonly used to describe text vectors.

The Vector Space Model (VSM) transfers the processing of text content to Vector operations of Vector Space[11], and we can express the semantic similarity by the similarity of Space. It is easy to understand and intuitive. When the document is represented as a vector in the document space, the similarity between the documents can be measured by calculating the similarity between the vectors[11]. The importance of keywords is reflected by their weight. Document D can be represented as $\{(k_1, w_1), (k_2, w_2), \ldots, (k_n, w_n)\}$, among them, k_i is the keyword of document D, w_i is the weight of k_i ($i=1, 2, \ldots, n$).

2. Feature extraction

If the dimensions of the text vector are represented directly by the word segmentation and the word frequency statistical method, the dimension of this vector will be very large. The untreated text vector not only brings the huge computational overhead to the follow-up work, the efficiency of the whole process is very low, and can damage the accuracy of classification and clustering algorithm so that it is difficult to get satisfactory results. Therefore, it is necessary to further purify the text vector and find out the most representative text features of the text feature category on the basis of guaranteeing the original meaning. In order to solve this problem, the most effective method is to reduce dimension by feature selection.

The main functions of the feature extraction are that decrease the number of words to be processed as far as possible without damage the main information of the text. In this way, we can reduce the dimension of vector space, so as to simplify the calculation, improve the speed and efficiency of text processing[12].

The methods of feature selection are:

(I) transform the original features into fewer new features by means of mapping or transformation;

(2) select some of the most representative features of the original features;

(3) select the most influential features according to the knowledge of experts;

(4) using the mathematical method, find out the most characteristics of classified information, this method is a more accurate method, human factors of interference are less, especially suitable for the application of text mining system.

Because the construction of the evaluation function is not particularly complex and the application scope is wide, more and more people prefer to use the construction evaluation function to select the features. At first, we can construct an evaluation function, evaluate each feature in the feature set, and score each feature so that each word gets an evaluation value, also known as weight. Then, all features are sorted by weight, and the optimal feature of the predetermined number is extracted as the characteristic subset of the extracted results. Obviously, for this type of algorithm, the main factor determining the effect of text feature extraction is to evaluate the quality of the function.

3. TF-IDF

The most effective way to implement the word weight is TF-IDF. The guiding ideology of TF- IDF is based on the basic assumption that words that appear many times in a text will appear many times in another similar text, and vice versa. Therefore, if the feature space coordinate system takes term frequency (TF) as the measure, it can reflect the characteristics of similar text, it also considers the difference between different categories of words. TF- IDF method introduce the concept of inverse document frequency (IDF), which means how many documents contain the term. Treat the product of TF and IDF as the feature space coordinate values measure. TF- IDF algorithm can be used to calculate the weight of key value that occurs in a document with higher frequency, while occurs in other documents with lower frequency. This method suggests that this kind of word has the stronger ability to make this document distinguished, its weight should be greater. [4]

To sort the weight of all words, there are two ways to choose according to the need:

(1) the maximum number of fixed number n keywords;

(2) the choice value is greater than the keyword of a certain threshold. In some

experiments for artificial selection features, 4- 7 is more appropriate.

4. User Modeling

Modeling is mainly based on user feedback, so user feedback has a critical impact on system performance. First, we need to consider how to get the data of the model, how to consider the changes in user interest and requirements, and how to model it. Among them, user information includes user's historical data, which can be used as input data of the model, and then the modeling method of TF-IDF and other user models is used for modeling.

5.**Similarity calculation**

While ranking the results of the search, we need to calculate the similarity between the patent and the search string. Similarity calculation is generally calculated by the distance between the feature vector. If the distance is small, the similarity is large; if the distance is large, the similarity is small.

Problem definition: there are two objects, X, and Y all contain N dimension features, $X = (x_1, x_2, x_3, ..., x_n)$, $Y = (y_1, y_2, y_3, ..., y_n)$, calculate the similarity between X and Y. The commonly used methods are as follows:

Euclidean Distance-based Similarity

Euclidean distance calculation similarity is the simplest and most understandable method in all similarity calculations. Calculate the absolute distance between points in a multidimensional space. When the data is dense and continuous, this is a good calculation. Because the calculation is based on the absolute values of the features of each dimension, the Euclidean measure needs to ensure that each dimension indicator is at the same scale level. See formula (2-1).

$$d(x,y) = \sqrt{(\sum (x_i\text{-}y_i)^2)} \qquad (2\text{-}1)$$

Cosine Similarity

Cosine similarity regards the cosine of the angle between two vectors as the measure of the difference between two vectors in vector space. Compared to distance metrics, the cosine similarity focuses more on the difference in direction of the two vectors, instead of the distance or length.

$$\sin\ (X,\mathrm{Y}) = \cos\theta = \frac{\vec{x}\vec{y}}{\|x\|\|y\|} \qquad (2\text{-}2)$$

According to the respective calculation methods and measurement characteristics of Euclidean distance and cosine similarity, they can be applied to different data analysis models respectively: Euclidean distance can reflect the absolute difference of individual numerical characteristics, while the cosine similarity is more suitable to distinguish the difference from the direction, but not

absolute value

Pearson correlation-based similarity

The Pearson correlation coefficient is generally used to calculate the closeness of the relationship between two fixed-range variables. Its value is between [-1, 1]. When the linear relationship between two variables increases, the correlation coefficient tends to be 1 or -1; when one variable increases, another variable also increases, indicating that they are positively correlated, the correlation coefficient is larger than 0; if a variable increases, another variable is reduced, indicating that they are negatively correlated, the correlation coefficient is less than 0; if the correlation coefficient is equal to 0, there is no linear relationship between them.

It is expressed by the mathematical formula that the Pearson correlation coefficient is equal to the covariance of the two variables divided by the standard deviation of the two variables. See the formula (2-3).

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E((X-\mu_X)(Y-\mu_Y))}{\sigma_X \sigma_Y} = \frac{E(XY)-E(X)E(Y)}{\sqrt{E(X^2)-E^2(X)}\sqrt{E(Y^2)-E^2(Y)}} \quad (2\text{-}3)$$

2.5 Expected goal

In this paper, we will design the search query optimization system based on the user's feedback information, generate the customized query suggestion for the existing patents alert, and rank the search results by calculating the similarity between patents and the new search string. In the ideal condition, the rank list of results obtained by the system can nearly match the actual feedback of the users, the irrelevant results are significantly reduced. It would have practical significance and could be integrated with the existing patent search system.

3 Schedule and work plan

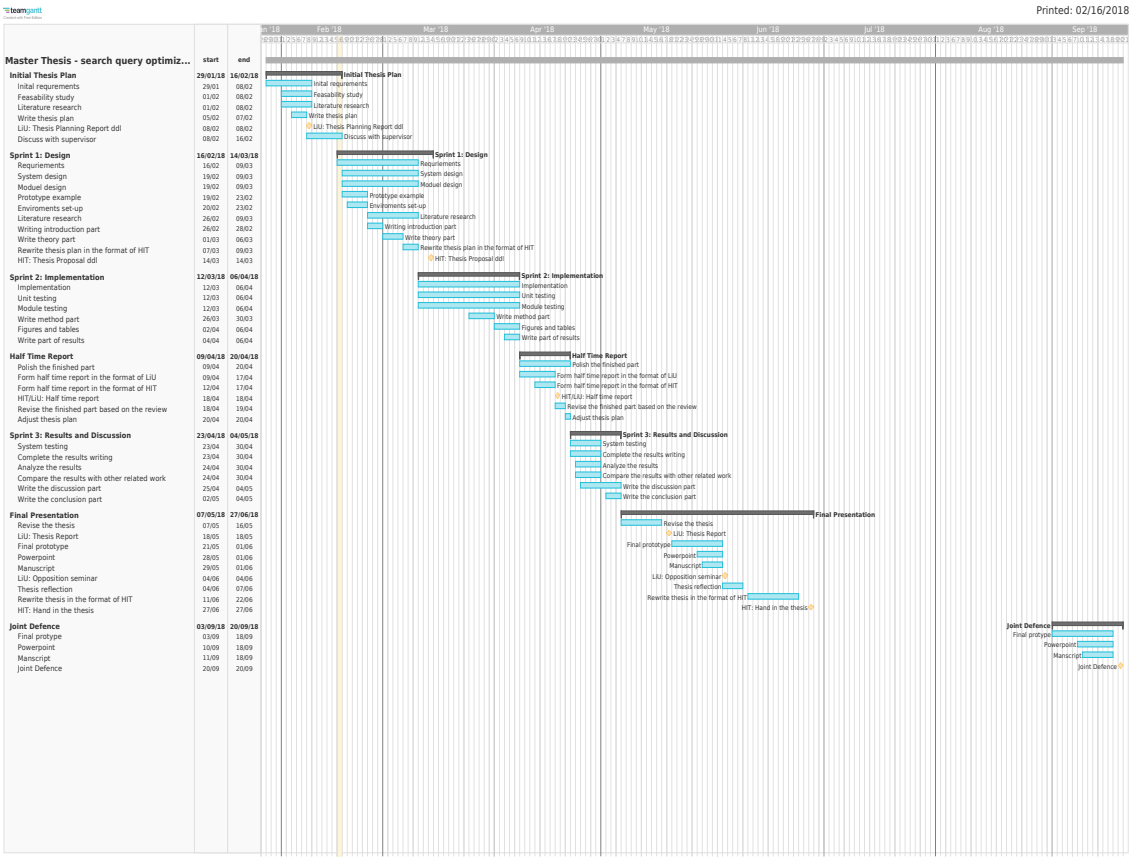
3.1 Schedule

Here is the time schedule for HIT-LiU joint thesis:

HIT - LiU Joint thesis time plan PRELIMINARY PLAN 171220	Minimum length LiU/HIT 20 Weeks LiU	 5 Months HIT	LiU HIT
Dates (example)			
2018-01-15	1		HIT: Internship Application deadline 22/1
2018-01-22	2		
2018-01-29	3		
2018-02-05	4		LiU: Thesis Planning Report Deadline 8/2
2018-02-12	5	1	
2018-02-19	6		
2018-02-26	7		
2018-03-05	8		
2018-03-12	9	2	HIT: Thesis Proposal Deadline 14/3
2018-03-19	10		
2018-03-26	11		
2018-04-02	12		
2018-04-09	13		
2018-04-16	14	3	HIT: Mid Term report "=" LiU Half time report Deadline 18/4
2018-04-23	15		
2018-04-30	16		
2018-05-07	17		
2018-05-14	18	4	LiU: Thesis Report to be defended at opposition seminar needs to be filed to supervisor and opponent. 18/5
2018-05-21	19		HIT: Defence verification (LiU/Tea Nygren will send BSc certificate and Transcript to Charlotte.)
2018-05-28	20		
2018-06-04			LiU: Opposition seminar 8/6. If the student also wants to do a separate LiU-defence - contact examiner @ LiU.
2018-06-11			
2018-06-18		5	HIT: Submit Internship Certification 27/6 Hand in the thesis for HIT supervisor thesis evaluation and feedback.
2018-06-25			LiU closed. HIT: The thesis will be updated until it is ready for defence approval
2018-07-02			LiU closed. HIT: The thesis will be updated until it is ready for defence approval
2018-07-09			LiU closed. HIT: The thesis will be updated until it is ready for defence approval
2018-07-16			LiU closed. HIT: The thesis will be updated until it is ready for defence approval
2018-07-23			LiU closed /HIT closed
2018-07-30			LiU closed /HIT closed
2018-08-06			LiU closed /HIT closed
2018-08-13			HIT closed
2018-08-20			HIT closed
2018-08-27			HIT closed
2018-09-03			
2018-09-10			
2018-09-17			HIT/LiU: Joint Defence. 20/9

3.2 Work plan and implementation

Here is a gantt chart that details my work plan:



Major references

[1] T. Ruotsalo, G. Jacucci, P. Myllymäki, and S. Kaski, “Interactive intent modeling,” *Commun. ACM*, vol. 58, no. 1, pp. 86–92, 2014.

[2] M. Allahyari *et al.*, “A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques,” 2017.

[3] J. Hu, S. Li, Y. Yao, L. Yu, G. Yang, and J. Hu, “Patent Keyword Extraction Algorithm Based on Distributed Representation for Patent Classification,” 2018.

[4] J. Li, K. Zhang, and others, “Keyword extraction based on tf/idf for Chinese news document,” *Wuhan Univ. J. Nat. Sci.*, vol. 12, no. 5, pp. 917–921, 2007.

[5] R. Wang, W. Liu, and C. McDonald, “Using word embeddings to enhance keyword identification for scientific publications,” in *Australasian Database Conference*, 2015, pp. 257–268.

[6] R. Mihalcea and P. Tarau, “TextRank: Bringing order into texts,” *Proc. EMNLP*, vol. 85, pp. 404–411, 2004.

[7] S. Don and D. Min, “Feature Selection for Automatic Categorization of Patent Documents,” *Indian J. Sci. Technol.*, vol. 9, no. 37, 2016.

[8] J. Joung and K. Kim, “Monitoring emerging technologies for technology planning using technical keyword based analysis from patent data,” *Technol. Forecast. Soc. Change*, vol. 114, pp. 281–292, 2017.

[9] C.-C. Wu, “Constructing a weighted keyword-based patent network approach to identify technological trends and evolution in a field of green energy: a case of biofuels,” *Qual. Quant.*, vol. 50, no. 1, pp. 213–235, 2016.

[10] R. C. Balabantaray, C. Sarma, and M. Jha, “Document Clustering using K-Means and K-Medoids,” *arXiv1502.07938 [cs]*, 2015.

[11] D. Chen, P. Zou, and N. Zhao, “A research on the construction process of integrated indicator system based on vector space model,” in *2014 IEEE International Conference on System Science and Engineering (ICSSE)*, 2014, pp. 201–204.

[12] R. Gongchang, L. Qi, and Y. Fenghai, “On Classification and Extraction of Deep Knowledge in Patents Based on TRIZ Theory,” in *2014 Fifth International Conference on Intelligent Systems Design and Engineering Applications*, 2014, pp. 666–670.

Evaluations of supervisors

Evaluation of Supervisor at Industry: Pass Supervisor (sign): F. He 2018-03-21 Y M D
Evaluation of Supervisor at Partner University: Supervisor (sign): Y M D
Evaluation of Supervisor at Harbin Institute of Technology: Supervisor (sign): Y M D
Evaluation about the proposal of the exam team: (sign): Y M D