

IMPROVING SEMANTICS OF QUERY BASED ON ONTOLOGY AND LOCAL ANALYSIS IN VIETNAMESE INFORMATION RETRIEVAL

Tuan Tran Anh^{*}

Received: November 05, 2015; Revised: February 05, 2016; Accepted: March 24, 2016

Abstract

The results from Information Retrieval (IR) having the same domain with a user's query are to decrease time in selecting useful information. However, IR sometimes returns the results which are not suitable for users. One of the reasons to affect the returned results of IR is the user's query without semantics. To improve these results of IR, the paper proposes a method combining ontology and local analysis technique to add semantics into original query.

Keywords: Vietnamese semantics information retrieval, Vietnamese language query expansion, ontology

Introduction

Nowadays, the World Wide Web (WWW) plays an important role in sharing information on the Internet, and it grows rapidly. The information on the Internet being shared, developed, and expanded by users is useful and exists mainly on web pages expressed in natural languages. To get useful information from the Internet, the search engine (SE) and Information Retrieval (IR) such as Google, Yahoo, Lucene, Solr, etc. have strongly been researched and developed. However, these systems return results which respond only a part of the users' requirements and have many

documents unsuitable for users. Furthermore, these systems just support well for languages, which detect a word by blank (Vietnamese language is not like that (Dinh, 2006)). This makes a huge impact on the precision and recall of relevant documents in Vietnamese language. The returned results of the IR and SE are not suitable for users because the main reasons are (Nguyen and Phan, 2008): these systems are not powerful enough to return expected results; the user's query has not fully reflected semantics to support the search engine and information retrieval. Therefore, a

Faculty of Foreign language – Informatics – Economics DakLak College of Pedagogy 349 Le Duan st, Buon Ma Thuot, DakLak, Vietnam. E-mail: tuanta@dlc.edu.vn

^{} Corresponding author*

semantics information retrieval system which is added semantics into the user's original query needs to be built in order to get the results as expected.

Related Works

There are many research groups on SEs and IRs. For English language, the CIRS (Concept-based Information Retrieval System) in Ngoc's research (Le, 2008) which is a semantics information retrieval system is based on MetaMap and XIOTA tools. The concepts of the CIRS are extracted from UMLS. In the research of Thanh and Tuoi (Nguyen and Phan, 2008; Nguyen, 2010), the Lucene - based IR completes a user's original query in a form of noun phrase and expands the query based on OOMP (Ontology of Object Member Property) ontology. Vuong and Tru (Ngo and Cao, 2009) propose ontology - based IR using a model which combines keywords and named entities (NE). The TREC dataset is used to test in this IR. For Vietnamese language, the search engine in Phuc and *et al.*'s research (Do *et al.*, nd) searches a word or phrase having close meanings to an original query and basing on ontology in a field of informatics. In their research, this engine uses the Oracle Text in order to index corpus and this feature is one of the advantages of this SE. The corpus size is larger because Vietnamese stopwords in corpus are not removed. Viet and *et al.* (Huynh *et al.*, 2010) develop the Vietnamese search system based on Lucene. This system supports two applications: an index on the server is created and updated by administrator in the first one; the last allows users to search in LAN and the Internet. In their work, the authors only refer tasks:

removing Vietnamese stoplist, normalizing, etc. in a document of the corpus in the indexing phase and do not refer to Vietnamese language processing as sentence and word segmentation in the indexing and searching phase. The VSearch developed by Thuan (Nhu, 2011) is Vietnamese search engine. VietSpider and VSearch Crawler which are two important parts of this engine extract web pages content. This engine does not use page rank technique and is based on Solr open source. For other language, the Thai search engine is called Sarsarn Look (Choochar, nd; NECTEC, 2010) based on Lucene and developed by NECTEC. The ThaiAnalyzer tool is built to pre-process Thai documents. This system has been developed with many different versions and components have been added such as ontology-based semantics search, Thai-English cross language information retrieval, etc. in each version.

Methodology

OOMP Ontology

OOMP is ontology and has hierarchy of semantics concepts based on relationships (R^m , R^p , $R^m_{\circ}$, $R^p_{\circ}$) (Nguyen and Phan, 2008; Nguyen, 2010). Figure 1 presents an OOMP ontology example.

The OOMP ontology structure is proposed by Thanh and Tuoi (Nguyen and Phan, 2008) and is constructed in relational data model (Nguyen and Phan, 2008; Nguyen, 2010). To support better ontology training in our model, this ontology is improved and presented as follow (Tran *et al.*, 2012):

- Object(oID, oValue, oPOS)
- Member(mID, mValue, mPOS, mPriority, mProbability)

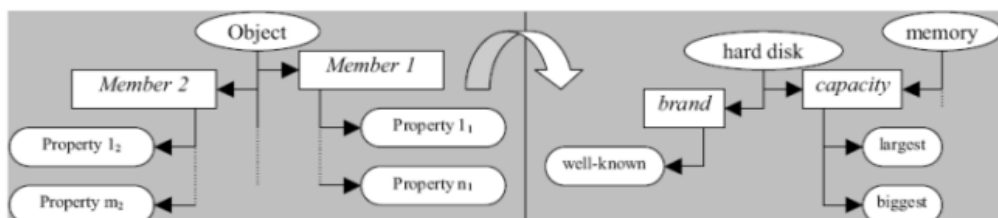


Figure 1. Example of OOMP ontology (Nguyen and Phan, 2008; Nguyen, 2010)

- Property(pID, pValue, **pPOS**, pPriority, pProbability)
- Rm(oID, mID, Dep)
- Rp(mID, pID, Dep)

where each element has some attributes as follows:

- oID, mID, and pID are the identification of element and primary key.
- Value is a content of element.
- Probability is used to save the probability $\wp(m_j)$ ($\wp(p_k)$) between Object and Member (between the Member and Property).
- Priority presents an element is the more dominant then the other is in a same condition.
- **POS** is pos tag of element such as Np, Nn, An, etc. This attribute which helps to determine POS tag of element easier is proposed by us.

Ontology Training

This method was adopted and adapted from Thanh and Tuoi's research (Nguyen and Phan, 2008). The three steps are as follows:

- Step 1: Extract candidate noun phrases

(candidate noun phrases are extracted from corpus in Vietnamese language). In this step, the documents will be extracted noun phrases by GATE and Jape built in Chau and Tuoi's research (Nguyen and Phan, 2008).

- Step 2: Determine the components O, M, and P basing on the pattern (Tran and Dao, 2013) from the noun phrase extracted in Step 1.

- Step 3: Build relations from O, M, P in Step 2, and then select the best results for ontology.

Vietnamese Thesaurus Dictionary Development

Vietnamese thesaurus dictionary called viSC is built by expert with the help of the Transact-MySQL and stored in tables in MySQL. The viSC is built as follows: Words (Lemma) in the computer field from WordNet are extracted. After being removed overlapping words, Synonyms with the words in the previous step are determined from WordNet, and removed duplicate words that are not

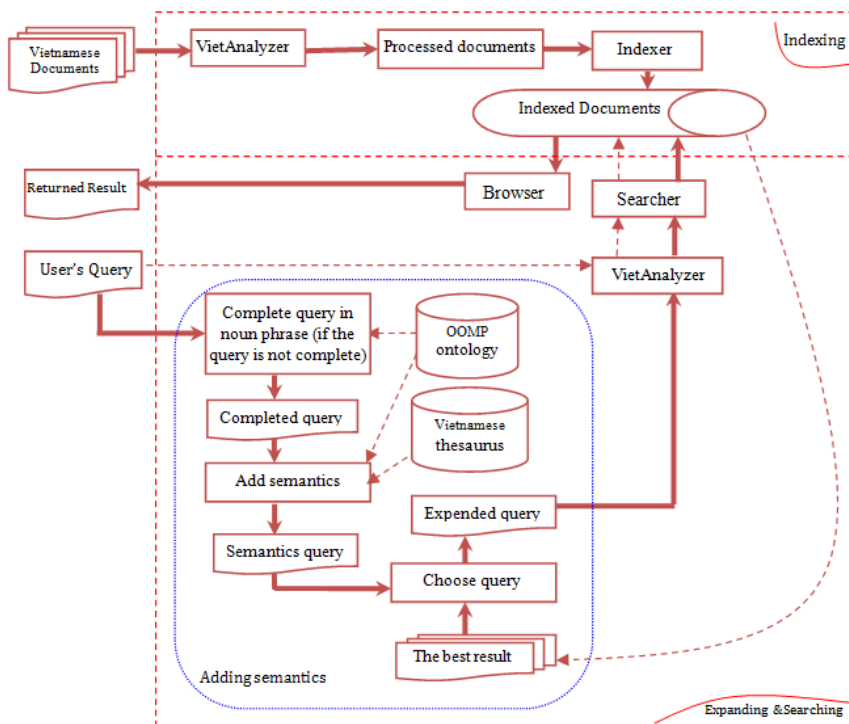


Figure 2. A proposed semantics model for Vietnamese Information Retrieval

related to selected areas. Then, these words are translated into Vietnamese via the Bambo online dictionary and LacViet dictionary. Next, the synonyms in the field of computer from the Vietnamese dictionary are selected and added to viSC dictionary. Last, abbreviations corresponding to the words in the viSC dictionary are added if any.

A Proposed Semantics Model for Vietnamese Information Retrieval

To add semantics into user's query for Vietnamese information retrieval, a model is proposed and illustrated in Figure 2.

There are two phases in the Vietnamese Information Retrieval. The first one is an indexing phase and the last is an expanding and searching phase.

In an indexing phase, a set of documents is normalized into a set of normalized documents by VietAnalyzer tool (Tran and Dao, 2013). After being normalized, these documents are input in the indexing phase. An indexing normalized Vietnamese document is a Vietnamese document which is normalized and is used for indexing phase. This document has a number of sentences which are in form of Unicode and do not contain Vietnamese stopwords.

In an expanding and searching phase, user's query is checked if it is complete or not. If the query is not complete, some components from the ontology are added into the query in order to complete in a form of noun phrase (*A query is complete if it consists description of the object, key member, key property and they are in R^m and R^p relations* (Nguyen and Phan, 2008; Nguyen, 2010). Next, semantics being added into query is called semantics query (the semantics query includes some noun phrases having the same domain). Then, semantics query is calculated the score based on the top ranked documents. After being calculated the score, noun phrases in the semantics query having high score will be selected and added into the user's query. The user's query is called the expanded query which has semantics and context with the user's query is converted to digital character by VietAnalyzer. After that, the query will be used to search documents related to query. The steps of the expanding and searching phase are detailed as follows:

• Step 1: Add semantics into user's query from ontology and thesaurus dictionary

Keywords in form of noun phrase extracted from ontology of OOMP will be called vnNP (Vietnamese Noun Phrase). To get better results with the Vietnamese language, the Similar Noun Phrase Expansion (SNPE) algorithm (Nguyen and Phan, 2008) is improved by using viSC thesaurus dictionary, then to produce *iSNPE* algorithm (improved Similar Noun Phrase Expansion). The *iSNPE* algorithm steps are detailed as follows (Tran *et al.*, 2012):

- **Step 1.1:** The user's query is verified if it is complete or not. If the query is not complete, the query will be added components from the ontology in order to complete the query in a form of noun phrase. This step is done by CNPV (Complete Noun Phrase Verification), NPC (Noun Phrase Completion) algorithm (Nguyen and Phan, 2008; Nguyen, 2010).

- **Step 1.2:** The completed query in step 1.1 will be analyzed into components (*Object – O, Member – M, and Property – P*). After being analyzed, synonyms corresponded the components (*O, M, P*) will be selected from Vietnamese thesaurus dictionary to add into sets of synonyms (*O_c, M_c, P_c*). The VnSS (*Vietnamese Synonym Selection*) algorithm (Tran *et al.*, 2012) does select synonym from thesaurus dictionary.

- **Step 1.3:** The similar noun phrases are generated from the sets of synonyms in step 2 by CNPG (*Complete Noun Phrase Generation*) algorithm (Nguyen and Phan, 2008).

• Step 2: Select the highest scored keywords and add into the user's query

- Documents retrieved and ranked by Lucene have context with the user's query. The top N documents, which are selected from the retrieved and ranked documents, are the best results. The N argument plays an important role in

calculating the score of the keywords.

- Calculate the score of the keywords in step 1 based on the best documents. The score is based on conditional statistics as equation belows:

$$\text{score}(\text{vnNP}) = \text{score}(t_1, t_2, \dots, t_h) \\ = \prod_{j=2..h} p(t_j | t_{j-1})$$

where: conditional probability of t_1 with t_2

- t_j is a word in the vnNP.
- $p(t_j | t_{j-1})$ is conditional statistics between two words

- Choose the M keywords have the highest score. The M argument also plays an important role in supporting IR, which returns results closer to the user's query.
- Add the M keywords into the user's query.

For example:

- With initial query "bộ nhớ máy tính" (computer memory).

• Step 1: Add semantics into user's query from ontology and thesaurus dictionary

- Query is tagged and resulted "bộ nhớ /Nc máy tính/Nc". The CNPV algorithm

is called and the retrieved result is $\{TH_5\}$.

- With $\{TH_5\}$, NPC is called to complete noun phrase and define $o_i = \text{"máy tính"}$, $m_j = \text{"bộ nhớ"}$ and $p_k = \text{"mới"}$.

- Basing on the viSC, we obtain $O_c = \{\text{"máy tính"}$, $\text{"máy vi tính"}\}$, $M_c = \{\text{"bộ nhớ"}$, "bộ lưu trữ" , $\text{"memory"}\}$, $P_c = \{\text{"mới"}$, $\text{"thấp"}\}$

- Basing on ontology, we retrieve the results: $\{\text{"bộ lưu trữ của máy tính mới"}$, $\text{"bộ lưu trữ máy tính mới"}$, $\text{"bộ lưu trữ máy vi tính mới"}$, $\text{"bộ nhớ của máy tính mới"}$, $\text{"bộ nhớ của máy tính thấp"}$, $\text{"bộ nhớ máy tính mới"}$, $\text{"bộ nhớ máy tính thấp"}$, $\text{"bộ nhớ máy vi tính mới"}$, $\text{"memory của máy tính mới"}$, $\text{"memory của máy tính thấp"}$, $\text{"memory của máy vi tính mới"}$, $\text{"memory máy tính mới"}$, $\text{"memory máy tính thấp"}$, $\text{"memory máy vi tính mới"}$, $\text{"memory máy vi tính thấp"}\}$

• Step 2: Select the highest scored keywords and add into the user's query

- Documents retrieved and ranked by Lucene have context with the user's query. Choose $N = 10$ and $N = 20$ in order to evaluate.

- With $N = 10$ and $N = 20$, the score of the keywords in step 1 is presented in the Table 1.

Table 1. The score of the keywords

No	Candidate noun phrase	Score	
		N = 10	N = 20
1	bộ lưu trữ của máy tính mới (new storage of computer)	0.047	0.056
2	bộ lưu trữ máy tính mới (new computer storage)	0.104	0.113
3	bộ lưu trữ máy vi tính mới (new computer storage)	0.043	0.041
4	bộ nhớ của máy tính mới (new memory of computer)	0.154	0.146
5	bộ nhớ của máy tính thấp (small memory of computer)	0.064	0.047
6	bộ nhớ máy tính mới (new computer memory)	0.261	0.290
7	bộ nhớ máy tính thấp (small computer memory)	0.109	0.093
8	bộ nhớ máy vi tính mới (new computer memory)	0.087	0.056
9	memory của máy tính mới (new memory of computer)	0.250	0.301
10	memory của máy tính thấp (small memory of computer)	0.104	0.097
11	memory của máy vi tính mới (new memory of computer)	0.104	0.110
12	memory máy tính mới (new computer memory)	0.778	0.729
13	memory máy tính thấp (small computer memory)	0.324	0.234
14	memory máy vi tính mới (new computer memory)	0.324	0.267
15	memory máy vi tính thấp (small computer memory)	0.135	0.086

- Choose the 10 keywords have the highest score. The result is presented in Table 2.

Results and Discussion

Data Source

For Vietnamese language, there currently is not a standard corpus like TREC. To evaluate experiment, a corpus in computer field must be built. The corpus is extracted from web (.html, .htm) to text (.txt) and normalized to Unicode. This corpus is divided into two parts: one is used to training ontology (D1, D2) and the other is used to indexing (D3, D4, D5, D6). The detail of corpus is presented in Table 3.

Experiment of a Proposed Semantics Model for Vietnamese IR

Lucene open source is used in our experiment; however, it does not support for Vietnamese language. To pre-process Vietnamese text for an information retrieval in indexing and searching, a package built to support is

called VietAnalyzer. In the model, two arguments M and N play an important role in evaluating experiment. The N argument is a number of the best ranked documents to calculate the score of the keywords and the M argument is a number of the highest scored noun phrases to add into the user's query. According to the research of Claudio and et al (Claudia *et al.*, 2001), the N which has the best result is 4 to 10; with N = 20, returned results are not good and decreasing; and the M argument is 10. So that, the experiment has 3 cases: the first is the proposed semantics model with N = 10, M = 10; the second is the proposed semantics model with N = 20, M = 10; and the last is ontology – based model. The experimental result is presented in Table 4.

As the result is presented in Table 4 and Figures 3 and 4, the proposed semantics model has precision higher than or equal to ontology – based model. In the experimental result, some cases are explained as follows:

- The first case, the proposed model's precision is greater than the ontology model's: An expanded query is to

Table 2. The 10 keywords have the highest score

N = 10; M=10		N = 20; M=10	
Candidate noun phrase	Score	Candidate noun phrase	Score
memory máy tính mới	0.778	memory máy tính mới	0.729
memory máy tính thấp	0.324	memory của máy tính mới	0.301
memory máy vi tính mới	0.324	bộ nhớ máy tính mới	0.290
bộ nhớ máy tính mới	0.261	memory máy vi tính mới	0.267
memory của máy tính mới	0.250	memory máy tính thấp	0.234
bộ nhớ của máy tính mới	0.154	bộ nhớ của máy tính mới	0.146
memory máy vi tính thấp	0.135	bộ lưu trữ máy tính mới	0.113
bộ nhớ máy tính thấp	0.109	memory của máy vi tính mới	0.110
bộ lưu trữ máy tính mới	0.104	memory của máy tính thấp	0.097
memory của máy tính thấp	0.104	bộ nhớ máy tính thấp	0.093

Table 3. Source of Data

Notation	Resources	Number of file	Capacity (MB)	Purpose
D1	Network administration	1571	111	Training
D2	PC World Magazine	1070	80.8	Training
D3	Vietnamese Wikipedia	92	5.8	Indexing
D4	Science and Technology	156	6.63	Indexing
D5	Information technology	660	46.0	Indexing
D6	IT News	609	34.2	Indexing

Table 4. The result of experiment

(1)	(2)	(3)	(4)	(5)	(6)
Q1	card đồ họa mới	a new graphics card	0.14	0.14	0.14
Q2	bộ nhớ máy tính	computer memory	1.19	1.19	0.20
Q3	màn hình tinh thể lỏng	liquid crystal displays	0.47	0.46	0.40
Q4	card mở rộng	expansion card	0.57	0.57	0.57
Q5	màn hình LCD	LCD screen	16.23	16.23	15.97
Q6	máy tính Dell	Dell Computer	0.20	0.20	0.20
Q7	máy chủ	server	47.80	47.80	47.18
Q8	máy tính	computer	99.52	99.52	97.28
Q9	bộ xử lý	processor	12.12	12.12	12.07
Q10	bộ nhớ	memory	62.70	62.70	36.21
Q11	màn hình	monitor	32.34	32.34	32.34
Q12	CPU AMD	CPU AMD	0.62	0.62	0.60
Q13	HDD box	HDD box	0.32	0.32	0.20
Q14	mainboard Gigabyte	Gigabyte mainboard	0.07	0.07	0.07
Q15	AMD	AMD	97.89	97.89	26.07
Q16	Asus	Asus	95.33	95.33	27.15
Q17	Samsung	Samsung	92.40	92.40	92.40
Q18	chipset	chipset	96.24	96.24	60.81
Q19	Intel	Intel	64.69	64.69	63.86
Q20	LAN	LAN	85.83	85.83	85.83
Q21	mainboard	mainboard	8.79	8.79	4.08
Q22	RAM	RAM	11.34	11.34	11.15
Q23	SATA	SATA	18.04	18.04	18.04
Q24	Seagate	Seagate	95.24	95.24	7.14
Q25	CRT	CRT	0.61	0.61	0.60
Average			37.63	37.63	25.62

Here:

(1): No., (2): Vietnamese query¹, (3): Vietnamese query in English

(4): Precision P (%) of model with N = 20, M = 10

(5): Precision P (%) of model with N = 10, M = 10

(6): Precision P (%) ontology – based model²

¹ <http://www.pcworld.com.vn/hoi-dap/phan-cung/>

² ontology - based model (Tran, 2015) is the same with the proposed semantics model, but it does not calculate score of keywords.

increase a number of retrieved documents and a number of relevant retrieved documents.

- The second case, the proposed model's precision is equal to the ontology model's: An expanded query and original query are the same.

The reason explains why experiment has result above: number of relevant documents is not still large and limit. This result shows that

a corpus also plays an important role in experimental evaluation. It is one of difficulties of the paper because a standard corpus for Vietnamese language does not construct at present.

Conclusions

This paper presents how to improve semantics of query in Vietnamese information retrieval.

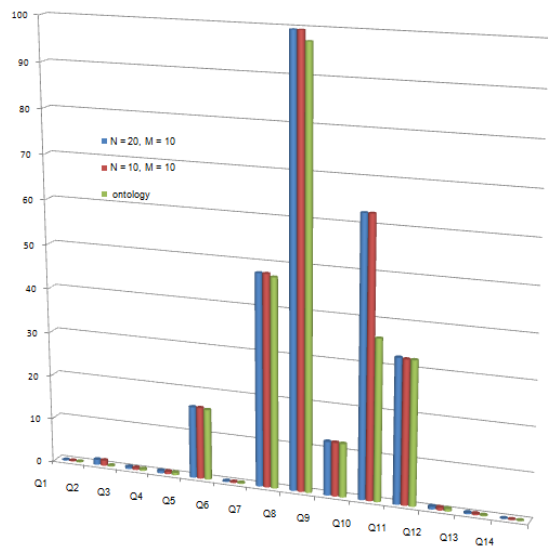


Figure 3. Statistic of Q1 to Q14 Precision

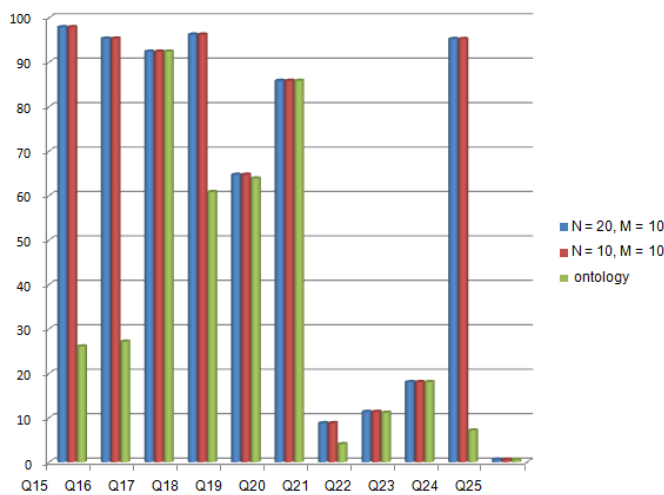


Figure 4. Statistic of Q15 to Q25 Precision

To do improve semantics of query, the proposed model is proposed. This model has two phases: the first phase determines noun phrases have semantics with the user's query from OOMP ontology; the last phase selects the highest scored noun phrases based on the best documents from corpus. In the experimental result, an average precision of the proposed semantics model (37.63%) is

greater than an average precision of the ontology - based model (25.62%). This result shows that the proposed model is better than the ontology - based model is. This Vietnamese information retrieval could use in field of library, agriculture, etc. Furthermore, this proposed semantics model could apply in some languages which do not detect a word by blank such as Thai, Lao, etc.

References

- Choochart, H. (n.d.). Information Retrieval and Search Engine. Research and Development on Information, National Electronics and Computer Technology Center.
- Carpineto, C., De Mori, R., Romano, G., and Bigi, B. (2001). An information-theoretic approach to automatic query expansion. *ACM Trans. Inform. Sys.*, 19:1-27.
- Fensel, D. (2000). The Semantic Web and its Languages, *IEEE Intell. Sys.*, 15(6):67-73.
- Dinh, D. (2006). Natural Language Processing textbook. Vietnamese National University of HCMC.
- Do, P., Do, H.C., Nguyen, T.T., Huynh, T.B.T., Nguyen, V.T., Nguyen, V.H., Nguyen, V.T., Pham, P.H., Duong, N.L.N., and Nguyen, P.T.H. (n.d). Develop a Search Engine supports for Information Retrieval in the field of Information Technology. HCMC University of Science.
- Huynh, D.V., Vo, D.T., and Vo, T.H. (2010). A case study on using open source lucene to build the full text search engine. *J. Sci. Technol.*, Da Nang University, 4(39):307-316.
- Le, T.N. (2008). Construct a Semantics Information Retrieval. Master thesis, HCMC University of Science.
- NECTEC. (2010). SanSarn Look Roadmap.
- Ngo, M.V. and Cao, H.T. (2009). A Generalized Vector Space Model for Ontology-Based Information Retrieval. *Vietnamese J. Inform. Technol. Commu.*, 22(2):43-53.
- Nguyen, C.T. (2010). Query Expansion Model Construction in Text Information Retrieval. Doctoral Dissertation, HCMC University of Technology.
- Nguyen, C.T. and Phan, T.T. (2008). Information Retrieval: An Ontology Solution for Query Complete. *J. ICT Minis. Inform. Commu.*, 19:84–92.
- Nguyen, Q.C. and Phan, T.T. (2008). Vietnamese Key Phrase Recognitions. *J. ICT Minis. Inform. Commu.*, 19:64–73.
- Nhu, D.T. (2011). Vsearch 1.0.
- Tran, A.T. (2015). Enhancing semantics of query in Vietnamese Information Retrieval. *Suranaree J. Sci. Technol* 22(2):197-204.
- Tran, A.T. and Dao, T.T.L. (2013). Vietnamese language information retrieval using open source. *J. Sci. Hanoi National Univ. Edu.*, 58(1):37–45.
- Tran, A.T., Phan, T.T., and Nguyen, C.T. (2012). Query expansion based on ontology for vietnamese query. *Proceedings of the 14th International Conference on Information Integration and Web-based Applications and Services (iiWAS2012)*, ACM; December 3-5, 2012; Bali, Indonesia, p. 332-335.