SEMANTIC-LITE RETRIEVAL ON IMPRECISE AND INCOMPLETE NATURAL QUERIES USING CONCEPTUAL GRAPHS

Thinh Nhat Phan[1], Tho Thanh Quan[1], Thien Cong Pham[1] and Nguyen Tuong Huynh[2]
[1] *University of Nguyen Tat Thanh, Hochiminh City, Vietnam*
[2] *Faculty of Computer Science and Engineering, Hochiminh City University of Technology, Hochiminh City, Vietnam*

Keywords: Natural query processing, Ontology, Conceptual graphs.

Abstract: Many attempts have been made on processing natural queries on relational databases, because this feature should render users virtually no boundary on expression search requests. Recently, research trends on this field have moved from compositional semantics into SQL-oriented processing on natural queries. However, this approach still suffers the problem of imprecise and incomplete information which often occurs when users submit their queries casually in practical situations. In this paper, we show an approach on translating the natural queries into conceptual graphs using semantic information captured by a domain ontology. Due to the well-structured representation of conceptual graph, we can resolve the impreciseness and retrieve the incompleteness when occurring. Experimental results have shown that our approach is quite promising.

1 INTRODUCTION

Relational databases have been used popularly in information systems due to the capability of precisely representing concepts and relations on certain domains. Moreover, due to their solid mathematical foundation, relational databases also allow user to query virtually all of desired information, commonly based on the well-defined Structured Query Language (SQL). However, for non-technical users, who happened to be the majority of actors interacting with most information systems, SQL is still hard-to-use due to its strict formal syntax and complicated structures needed to express the query goals.

Naturally, many attempts have been made to support users to query over relational databases by natural language forms. Early works have focused on compositional semantics (Androutsopoulos, Ritchie and Thanisch, 1995). However, this approach failed to deal with quantified queries. More recent works investigated translating natural queries into SQL (Frost and Fortier, 2007). Even though this approach can handle deep semantics in some queries, it required users to submit the queries in the forms whose syntaxes should be "similar" to that of the SQL. Nevertheless, if the query is rephrased as another form, this approach suffered much difficulty to process. We consider this problem as *imprecise information*.

With the recent advancement of the Semantic Web (Berners-Lee, Hendler and Lassila, 2001), *ontology* is considered highly useful to extract semantics from natural queries (Storey, Sugumaran and Burton-Jones, 2004). Remarkably, a work on ontology-based lightweight natural processing is reported (Kang, Na, Lee and Yang, 2004), whose method is able to process semantic-enriched queries. However, in practical situations, users are likely to submit queries in casually shorterned forms. Due to the missing of some important "anchor words", such shortened forms are hard to be processed efficiently. We consider this problem as *incomplete information*. Recently, to process natural queries, conceptual graph (CG) (Sowa, 1997) is regarded as an effective technique to capture and represent the semantics in linguistic structures. An effort on supporting retrieval using natural query over CG-represented documents was reported by Shady, Karray and Kamel (2007), in which the authors adopted some manual methods. However, automatic translation of natural queries into the corresponding CGs is still a complex and challenging problem. In this paper, we propose a new approach on natural query retrieval based on CG using domain ontology, which can

Nhat Phan T., Thanh Quan T., Cong Pham T. and Tuong Huynh N..
SEMANTIC-LITE RETRIEVAL ON IMPRECISE AND INCOMPLETE NATURAL QUERIES USING CONCEPTUAL GRAPHS.
DOI: 10.5220/0003501202630268
In *Proceedings of the 6th International Conference on Software and Database Technologies* (ICSOFT-2011), pages 263-268
ISBN: 978-989-8425-76-8

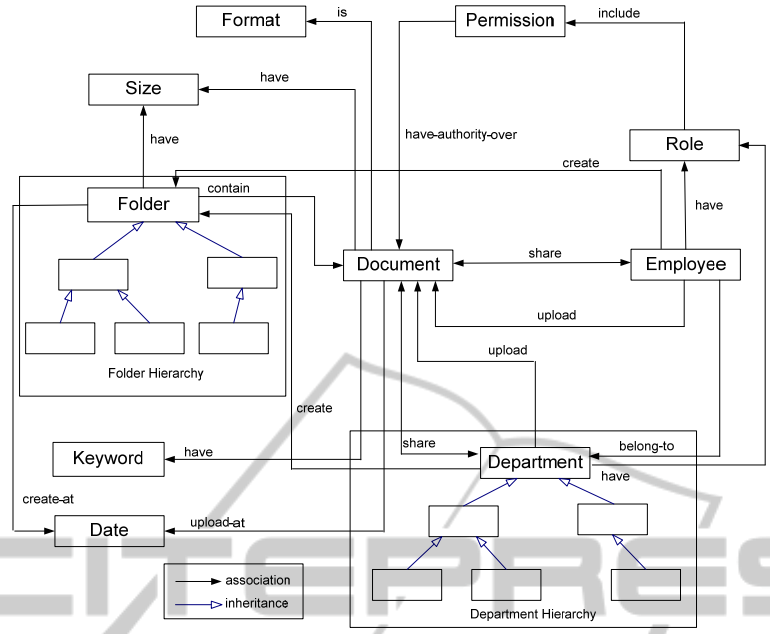


Figure 1: Conceptual Schema of the Document Management Ontology.

handle flexibly natural languages in freely casual forms. Figure 1 gives the conceptual schema of the Document Management Ontology, which will be used througout the paper. The Document Management Ontology has three concepts (or classes): *Document*, *Employee* and *Department*. The attributes of the class *Document* are major properties to identify the documents such as *keyword*, *format*, *size*, etc. The *Employee* class represents a set of staff that is granted to access to some resources according to their roles. The *Department* concept refers to the offices of the company, which are organized to manage each other.

The rest of this paper is organized as follows. Section 2 discusses the basic concept of CG and its represention for natural query. Section 3 presents our proposed approach for generating conceptual graph from natural query using domain ontology, which can be constructed automatically from an enhanced relational database. Section 4 gives some initial experimental results. Finally, Section 5 concludes the paper and discusses the direction for our future research.

2 QUERY REPRESENTATION USING CONCEPTUAL GRAPH

A *conceptual graph* (CG) is a notation for logic based on existential graphs. A CG can be displayed as a directed graph whose nodes can be a concept node or relation node. A concept node implies an individual of a concept, while a relation node indicates the relationships between individuals. Figure 2 shows an example of a CG, whose concept nodes are presented in boxes and relation nodes in ovals. The fact conveyed by this CG is that the author Peter has uploaded the report entitled "*The financial statement in 2010*". In this CG, the individuals Peter and "*The financial statement in 2010*" are called *individual referents* of the concepts Employee and Document respectively.

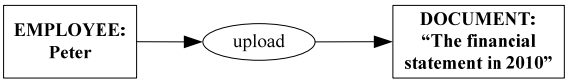


Figure 2: An example conceptual graph.

Figure 3 shows a CG representing the query "*Find the documents were uploaded by the employees of sales department of pdf format*" Notice that the "?" symbol indicates a *query referent* of the CG, which specifies the object that will be searched. The "*" symbol indicates a *generic referent*, which means that no specific individual of the concept *Employee* is mentioned explicitly in the query.

3 AUTOMATIC GENERATION OF CG-BASED QUERIES

In this section, we present an approach for automatic

generation of CG-based query from natural query using the knowledge captured from domain ontology.
The automatic generation process consists of the following two steps *Concept Generation* and *Relation Construction*.

3.1 Concept Generation

This step aims to parse the submitted query in order to identify the ontological concepts and individuals. For example, let us consider the following query:

(Q1):"Find all the documents were uploaded by employee named David."

Using the Document Management Ontology given in Figure 1, we can recognize the ontological concepts Document, Employee and the ontological individual "David" of the ontological class Employee from the query.

After recognizing ontological concepts and individuals in a query, we will map them into the corresponding concept nodes in the final CG. To do this, the following heuristic rules are used:

- An ontological individual will be mapped as an individual referent.
- An ontological concept will be mapped as a query referent if there is no individual of this concept recognized in the query.

Therefore, for the query (Q1), the individual of *"David"* will be mapped as an individual referent. Between the two ontological concepts *Document* and *Employee*, we only preserve the concept *Document* as a query referent since the concept *Employee* has the corresponding individual (i.e. *"David"*) found in the query.

Handling Imprecise Information. Note that when constructing domain ontology, the ontology engineer can define an ontology vocabulary that helps to recognize different keywords that may be relevant to certain ontological concepts and individuals. For example, the ontology engineer may define that the concept *Document* may be referred to by some linguistic terms such as *document*, *announcement*, *article*, *report*, *file*, etc. Thus, if the natural query is changed to another form like "Which the files David was uploaded?" the same ontological concepts and individuals will be recognized.

3.2 Relation Construction

In this step, we construct the relations between these concept nodes. First, for each pair of query referent and individual referent identified, we find a path between the corresponding ontological concepts in the conceptual schema of the domain ontology. Then, we unify all the paths to form the final CG-based query.

For example, in query (Q1), there is one query referent of *Document* and one individual referent of *Employee* identified. The corresponding path between these concepts in the Document Management Ontology is given in Figure 4, which is also the final CG-based query generated.

Figure 4: CG-based query generated from query (Q1).

Handling Incomplete Information. Consider the following query:

(Q3):"What did the employees of sales department upload at the month September 2010?"

There are two individual referents Department and *Date* identified. Figure 5 shows the path between these concepts in the Document Management Ontology, and the corresponding CG-based query. Note that in the identified path, an additional concept *Document*, which is mapped as a query referent in the CG, is also created. In other words, incomplete information of "*Document*" concept is automatic detected and retrieved when the CG is constructed.

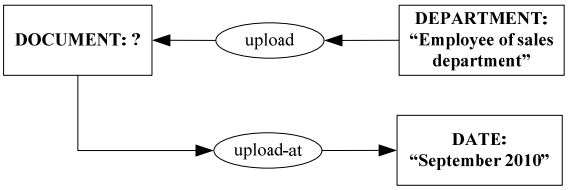


Figure 5: CG-based query generated from query (Q3).

3.3 Uncertainty Resolution

When parsing natural query, it is sometimes difficult to be certain about the concepts and individuals associated with a keyword, even when the ontology vocabulary is used. Here, we discuss some strategies that can be used to handle such uncertainties.

Multiple Individuals of the Same Ontological Concept. In such case, we can just simply generate a single node corresponding to these individuals. For example, in the following query:

(Q4):"Find documents are doc or docx format."

There are two individuals *doc* and *docx* of the same concept *format* identified. Figure 6 shows the final CG-based query.

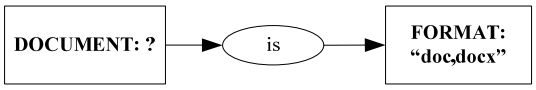


Figure 6: The CG-based query for query (Q4).

Multiple Concepts of the Same Keyword/Phrase. Our solution to this problem is to consider all the generated CG-based queries. For example, in the following query:

(Q4):"Find documents were stored in sales."

There are two CG-based queries generated as shown in Figure 7. Obviously, the results obtained when processing these queries should be relevant to the request from users.

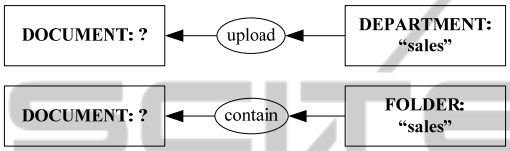
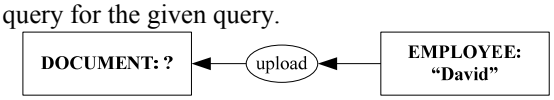


Figure 7: The CG-based queries for query (Q4).

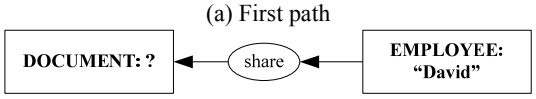
Multiple Paths of the Same Pair of Query/Individual Referents. In such case, we will select the most reasonable path whose relation nodes are most similar to the query keywords. For example, in the following query:

(Q5):"Find reports of employee David."

In the Document Management Ontology, there are two possible paths between the concepts *Document* and *Employee*, as depicted in Figure 8a and Figure 8b. However, the term *of* in the query should be relevant to the ontology vocabulary defined for the ontological relation *upload* in the Document Management Ontology. Thus, the first path, i.e. the path depicted in Figure 8a, should prevail and be selected as the appropriate CG-based query for the given query.



(a) First path



(b) Second path

Figure 8: Two paths identified for query (Q5).

4 EXPERIMENTAL RESULTS

In this section, we present an initial experiment to evaluate the effectiveness of the proposed ontology-based approach for natural query retrieval. In the experiment, we construct a simple Document Management System from the working information system of the EVietSoft company, a professional software development company (http:// www.evietsoft.net/GCMS.aspx).

The company has some the departments follows director's office, business department, technical department, accounting department, etc. The director department has the most powerful and controls all the others of company. Similarly, the senior departments will manage the junior department base on roles and permissions. We developed a plug-in into the current system allowing users to submit natural queries according to their working requirement. Table 1 gives the most common natural queries submitted by the company staff.

We have also compared the performance of our approach with other typical information retrieval (IR) techniques in terms of typical IR measures like recall, precision and F-measure, using the same query set as depicted in Table 1. Two techniques are used for comparison. The first technique applies the typical *tfidf* vector-space-model (VSM) to retrieve information from the input queries. In the second technique, we first cluster the data before applying the VSM for retrieval. Since the data are multi-dimensional and the queries are multi-objective, the multi-clustering technique (Quan, Hui and Fong, 2003) is adopted here. Table 2 presents the performance comparison of the different techniques based on recall, precision and F-measure.

As can be seen in Table 2, when combined with the multi-clustering technique, the performance of the VSM-based retrieval technique has been improved significantly in terms of both recall and precision. It is because when data are clustered using the multi-clustering technique, information can be represented better in clusters, thereby enhancing the retrieval performance. As compared to the VSM-based multi-clustering technique, the precision obtained by the CG-based query processing method is slightly lower. However, the recall is better since if the CG-based queries are generated precisely, the retrieval performance can achieve with almost absolute accuracy. As a result, the CG-based query processing technique has achieved the best performance in terms of the F-measure.

5 CONCLUSIONS

This paper has proposed an ontology-based approach for natural query retrieval using conceptual graphs. We have applied the proposed approach for

Table 1: Examples of common natural queries.

1. Find all documents of Peter employee
2. Find all documents are pdf format
3. Show me some reports of sales department contain the keywords "financial statement"
4. Documents of accounting department were uploaded in 2010
5. Search any documents of sales department are "excel 2003" and pdf format
6. I want to find any documents that I can download
7. Look up the documents are "word 2007" format in my Announcements folder
8. Look up all of documents that people has staff role can download
9. I want to read the software instruction manuals of technical department
10. Find the folders were created on 06-18-2010
11. Search all documents are less than "1 Mb" and are word or pdf format
12. Which the documents were created in "February 2010" that the employees of financial department can be deleted and viewed?
13. Which the documents are pdf format that Peter shares with me in Public folder?
14. The employees have the manageable role can download or delete the documents
15. List all of the documents were uploaded from 01-20-2010 to 09-20-2010 and are excel format
16. Search any documents in Report folders have the size from "500 Kb" to "1.5 Mb"
17. James was sent the announcements to sales department in October?
18. I want to download the documents of sales, accounting and financial department
19. List the folders I have created after "Jan 2010"
20. Show me the documents that contain the keywords "computer science" belong to "Plans" folder
21. Find all of the folders have the size are less than "5 Mb" and are created in September 2010
22. Find the documents of technological department were uploaded at 11-16-2010 contain the keyword "ontology"
23. The head of sales department shares the documents between everyone works for accounting department
24. Which the documents technological department can be download from "May 2009" to "May 2010"
25. Which the documents employees of financial department can be view or share in "Financial Department's Public Documents" folder
26. Find any documents are .doc or .docx format and its sizes are more than "1.2 Mb"
27. Which the folders sales department was created from "January 2010" to "December 2010" and more than "8 Mb"
28. David employee uploaded the documents contains the keywords "conceptual graph"
29. Which the documents of sales department that the head of financial department can be view?
30. The documents are docx format, upload from 01-10-2010 to 10-10-2010 and Peter share it with me

Table 2: Performance comparison on retrieval.

Techniques	Recall	Precision	F-measure
Vector-space-model (VSM)	78%	87%	82%
VSM + Multi-clustering	92%	94%	93%
CG-based Queries	98%	93%	95%

the retrieval of information in document management system. The initial experimental results have shown that our proposed approach is capable of handling effectively most of the typical search requests in natural language. By avoiding using a fixed grammar and making use of a domain ontology, our approach can handle the problem of *imprecise* and *incomplete* information. In addition, minor grammatical errors, which may probably occur in queries submitted casually by users in many practical situations, can also be tolerated reasonably.

ACKNOWLEDGEMENTS

This research project is funded by University of Nguyen Tat Thanh, Ho Chi Minh City, Vietnam We are also grateful for the technical helps of the EViet software company in terms of hosting services and experimental data provided.

REFERENCES

Androutsopoulos, I., Ritchie, G. D. and Thanisch, P. (1995). Natural Language Interfaces to Databases: An Introduction. *Journal of Natural Language Engineering*, 1(1), 29-81.

Berners-Lee, T., Hendler, J. and Lassila, O. (2001). The Semantic Web. *Scientific American*. Retrieved May 17, 2001 from http://www.scientificamerican.com/article.cfm?id=the-semantic-web

Cimiano, P., Haase, P., Sure, Y., Völker, J. and Wang, Y. (2006). Question answering on top of the BT digital library. *ACM Publisher*, In *Proceedings of the 15th International Conference on World Wide Web*, 861-862. doi:10.1145/1135777.1135915

Frost, A. R. and Fortier R. J. (2007). An Efficient Denotational Semantics for Natural Language Database Queries. *Springer-Verlag*, In *Proceedings of the 12th International Conference on Applications of Natural Language to Information*, 4592, 12-24. doi:10.1007/978-3-540-73351-5_2

Guarino, N. and Giaretta, P. (1995). Ontologies and Knowledge Bases - Towards a Terminological Clarification. *IOS Press, Toward Very Large*

Knowledge Bases: Knowledge Building and Knowledge Sharing, 25-32.

IntraText Digital Library, Available at: http://www.intratext.com/CERCA/Aiuto.htm.

Kang, I. S., Na, S. H., Lee, J. H. and Yang, G. (2004). Lightweight Natural Language Database Interfaces. *Springer-Verlag,* In *Proceedings of the 9th International Conference on Applications of Natural Language to Information*, 3136, 167-187. doi:10.1007/978-3-540-27779-8_7

Kim, S. S., Myaeng, S. H. and Yoo, J. M. (2005). A Hybrid Information Retrieval Model Using Metadata and Text. *Springer-Verlag,* In *Proceedings of the 8th International Asian Digital Libraries Conference*. 3815, 232-241. doi:10.1007/11599517_27

Lim, E. P. and Sun, A. (2006). Web Mining - The Ontology Approach. *In Proceedings of the International Advanced Digital Library Conference*.

Pham, T. D. (2000). Computing with Words in Formal Methods. *John Wiley & Sons*, *International Journal of Intelligent Systems*, 15(8), 801-810. doi:10.1002/1098-111X(200008)15:8<801::AID-INT7>3.0.CO;2-Z

Quan, T. T., Hui, S. C., Fong, A. C. M. and Cao, T. H. (2006). Automatic fuzzy ontology generation for Semantic Web. *IEEE Transactions on Knowledge and Data Engineering*, 18(6), 842-856. doi:10.1109/TKDE.2006.87

Quan, T. T., Hui, S. C. and Fong, A. C. M. (2003). Mining multiple clustering data for knowledge discovery. *Springer-Verlag,* In *Proceedings of the Discovery Science Conference*, 2843, 452-459. doi:10.1007/978-3-540-39644-4_45

Saracevic, T. and Dalbello, M. (2001). A Survey of Digital Library Education. In *Proceedings of the American Society for Information Science and Technology*, 38, 209-223.

Shady, S., Karray, F. and Kamel, M. (2007). Enhancing text retrieval performance using conceptual ontological graph. In *Proceedings of the 6th IEEE International Conference on Data Mining*, 39-44. doi:10.1109/ICDMW.2006.71

Sowa, J. F. (1997). Matching logical structure to linguistic structure. Houser N., Roberts D.D. and Van EvraJ. (Eds): *Studies in the Logic of Charles Sanders Peirce*, *Indiana University Press*, 418-444.

Storey, V. C., Sugumaran V. and Burton-Jones A. (2004). The Role of User Profiles in Context-Aware Query Processing for the Semantic Web. *Springer-Verlag,* In *Proceedings of the 9th International Conference on Applications of Natural Language to Information*, 3136, 45-62. doi:10.1007/978-3-540-27779-8_5

Varadarajan, R. and Hristidis, V. (2006). A system for query-specific document summarization. *ACM Publisher,* In *Proceedings of the 15th ACM international conference on Information and knowledge management*, 622-631. doi:10.1145/1183614.1183703