

A Topic Maps-based ontology IR system versus Clustering-based IR System: A Comparative Study in Security Domain

Myongho Yi¹ and Sam Gyun Oh^{2*}

¹ School of Library and Information Studies,
Texas Woman's University, P.O. Box 425438, Denton, TX 76204-5438
topicmap@gmail.com²

Department of Library and Information Science, Sungkyunkwan University, Myongryun-Dong
3-53, Jongro-Gu, Seoul, Republic of Korea
samoh@skku.edu

Abstract. Most clustering methods for information retrieval application do not work efficiently when dealing with complicated data. In this paper, we compare the performance of the Topic Maps-based method with the Clustering-based method. An experimental test was carried out using 20 volunteer to evaluate and compare the performance of the Topic Maps-based Information Retrieval system and Clustering-based Information Retrieval system in security domain. The experimental results show that a Topic Maps-based method provides both better recall/precision and shorter search time/search steps.

Keywords: Clustering, Ontology, Recall, Search Time, Topic Maps

1 Introduction

Many information organization approaches such as taxonomy, thesaurus, classification, and ontology have been attempted to provide effective searching. Among them, clustering and ontology approaches have received much attention. However, there have not been many studies which compare in terms of user performance. Previous studies have been attempted to demonstrate the benefits of clustering or ontology. Therefore, the comparison of each Topic Maps-based and clustering-based approach seemed valuable.

The purpose of this study is to compare the performance of our Topic Maps-based method with the Clustering-based method. In order to measure performance, this study implements a Topic Maps-based Security Information Retrieval (TMIR) system and Clustering-based Security Information Retrieval (CIR) system. Recall/Precision, search steps taken, and search time spent for given tasks are measured.

This paper has been organized as follows: section 1.1 will discuss research questions. Section 2 will discuss related works. Section 3 will introduce security domain. Section 4 will describe the development of TMIR and CIR. Section 5 will discuss research design. Section 6 will present the test results. Section 7 will conclude the paper.

2 Research Questions

The purpose of this study is to determine how the Topic Maps-based information retrieval system differs from clustering-based information retrieval system. The study poses the following research questions.

- 1 Are there recall/precision differences between TMIR and CIR?
- 2 Are there search time differences between TMIR and CIR?
- 3 Are there search steps differences between TMIR and CIR?

3 Related Works

Clustering is the classification of data into different subtopic categories. Clustering shows related items according to their similarity. Numerous clustering algorithms have been studied (E.K.F. Dang, Luk, Ho, Chan, & Lee, 2008; Nosovskiy, Liu, & Sourina, 2008). Two main approaches for clustering methods that have been used for data clustering in information retrieval are partitioning and hierarchical (E.K.F. Dang, et al., 2008). One of the limitations of clustering-based approach is that the relationships between terms are still implicit and require prior knowledge to make a relevance judgment. In other words, clustering-based search engines provide related terms by various algorithms; it shows gaps between clustered and user's categories.

Cluster-based search engines differ from ontology. While ontology explicitly reveals equivalence, hierarchical, and associative relationships to the user, cluster-based search engines only show related terms. For example, a user's

search for “security” using a cluster-based search engine retrieves “homeland security,” “security services,” “security resources,” etc. Ontology shows various relationships including related terms; therefore, the user’s judgment about relevance is better supported. One way to minimize the gap between system and user is to add rich semantic relationships among terms.

While clustering attempts semantic clustering, there is an absence of evidence about which semantic clustering can enhance searching. While many researchers address the potential of clustering (Biren Shah, Raghavan, Dhatric, & Zhao, 2006; Dunlavy, O’Leary, Conroy, & Schlesinger, 2007; Hu, Zhou, Guan, & Hu, 2008; Lin, Li, Chen, & Liu, 2007; Liz Price & Thelwall, 2005; Na, Kang, & Lee, 2007; Niall Rooney, Patterson, Galushka, Dobrynin, & Smirnova, 2008; Oscar Loureiro & Siegelmann, 2005; Ronald N. Kostoff & Block, 2005; Sherry Koshman, Spink, & Jansen, 2006; VicencTorra, Lanau, & Miyamoto, 2006), automatic clustering (Nosovskiy, et al., 2008) , cluster-based data mining (Busygin, Prokopyev, & Pardalos, 2008), and cluster-based information retrieval (Kang, Na, Kim, & Lee, 2007), there are few studies that compare user performance with ontology.

Topic maps are one of the two standards ontology languages. Using Topic Maps, users can browse rich semantic relationships among data. Unlike the clustering-based approach, semantic relationships are explicitly shown to users. The assignment of relationship labels to terms can be done automatically. Many structured resources such as metadata, XML, and database schemes contain information that can be automatically converted to terms, term types, and associations. Topic maps have topics that represent subject or terms (Garshol, 2002). Associations are used for linking among topics. An occurrence has actual resources that linked to topics.

4 Security Domain

For this study, security domain was chosen. Security is a complicated domain. Based on information security certification organization, security can be classified into ten domains (ISC, 2008).

- Access Control Systems and Methodology
- Telecommunications and Network Security
- Application and Systems Development Security
- Cryptography
- Security Management Practices

- Computer Operations Security
- Security Architecture and Models
- Law, Investigation, and Ethics
- Business Continuity Planning and Disaster Recovery Planning
- Physical Security

These ten domains can be classified into three broader categories. The first five domains belong to technical security, the next four domains belong to managerial security and the last domain belongs to physical security.

5 Development of Topic Maps-based Information Retrieval (TMIR) and Clustering-based Information Retrieval (CIR)

Development of TMIR and CIR system involves several steps and Figure 1 shows the processes to develop TMIR and CIR systems.

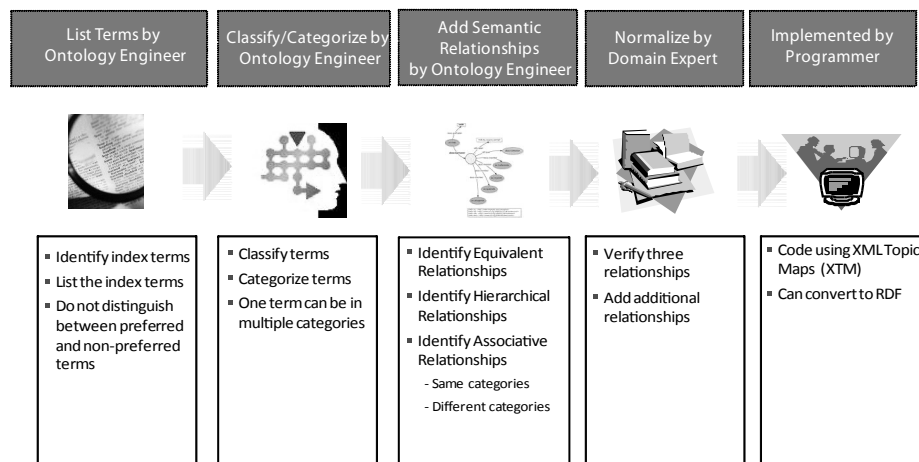


Fig. 1. Development Process of the TMIR and CIR System

5.1 List Terms

The terms are based on the search results from the one of the leading clustering-based search engine, clusty.com (<http://www.clusty.com>). The search with the

term “security” was conducted using clusty.com on June 7, 2008 and the total number of results is 249. The results are classified into two levels (see Table 1). The first level has 48 categories and the second level has 72 categories. The total number of unclassified or so called “Other Topics” is 20.

5.2 Ontology Modeling

Listed terms are converted to Topic Maps-based approach. Table 1 shows two different relationships: clustering-based relationships and Topic Maps-based Relationships. For example, network security (1.1) in clustering-based relationships is classified under information security category and the same network security (1.1) can be labeled as telecommunication, network and Internet security (1.1) in Topic Maps-based approach. The reclassification of clustering-based approach results in well-structured security domain.

Table 1. Data Clustering Ontology

Clustering-based Approach	Topic Maps-based Approach
1. Information Security (16) 1.1. Network Security (3) 1.2. Customers (2) 1.3. Valuable (2) 1.4. PGP (2) 1.5. Other Topics (7) X	1 Equivalent to Security 1.1 Tele., Network & Internet 1.2 Organization/Clients 1.3 1.4 Resources/Standard 1.5
2. Gov (14) 2.1. Social Security Administration (2) 2.2. Department (2) 2.3. Security Police (2) 2.4. Computer Security Resources (2) 2.5. Other Topics (6)	2 Organization 2.1 Organization/Government 2.2 Organization/Government 2.3 Organization/Government 2.4 Resources/Website
3. Alarm (17) 3.1. Monitoring (4)	3 Product/Hardware 3.1 Product/Monitoring

3.2. Security guards (3) 3.3. Equipment, Systems, Surveillance (3) 3.4. Automation (2) 3.5. Intercoms And Access Control Systems (2) 3.6. Focusing, CCTV (2) 3.7. Other Topics (2)	3.2 Person/Specialty 3.3 Product/Hardware 3.4 Product/Hardware 3.5 Product/Hardware 3.6 Product/Hardware 3.7
4. Bank (15) 4.1. Security Exchange (2) 4.2. Savings Bank (2) 4.3. Security State Bank (2) 4.4. Other Topics (9)	4 Organization 4.1 4.2 4.3 4.4
5. Homeland Security (8) 5.1. Department of Homeland Security (3) 5.2. Discussion Forums (2) 5.3. Other Topics (3)	5 Organization 5.1 Organization/Government 5.2 Resources/Dis. Forums 5.3

Ontology modeling is the next step after identifying the data. The ontology modeling process involves building relationships among data. Security ontology modeling is displayed in Figure 2 below.

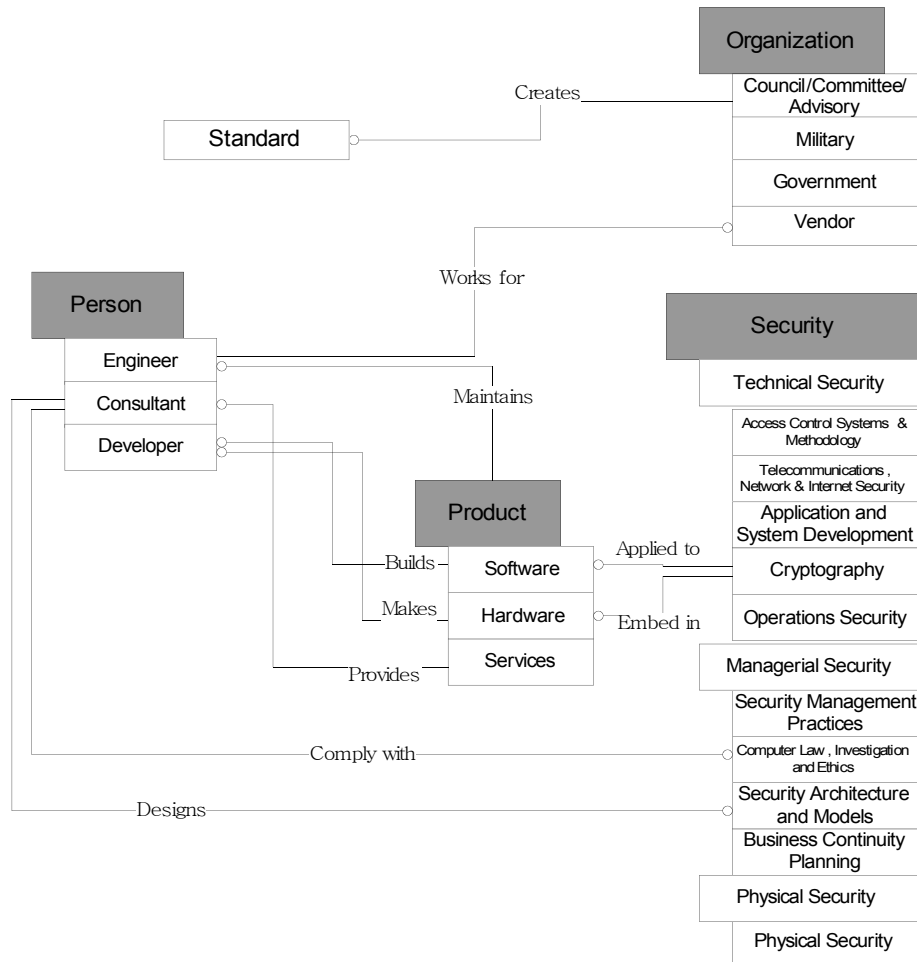


Fig. 2. Security Ontology

Figure 2 shows relationships among terms, and these relationships provide seamless connections among terms. More specifically, when a user searches for software package with specific cryptographic algorithm such as RSA, relationships in Figure 2 allow user to navigate different cryptographic algorithms and related products. The useful association for this search query will be “Applied to” between cryptography and software. Another example for security ontology is as follows: When a user wants to find out an engineer who maintains firewall. A user can find relationships between product and person very easily by browsing relationships between engineer and product.

5.3 Implementation of TMIR and CIR Systems

In order to examine user performance, a TMIR and a CIR system were implemented and a comparative experiment was conducted in which the performance of a TMIR system was compared to that of a CIR system. User performance for both systems was compared and contrasted using an experimental retrieval test, with the only difference between a TMIR and a CIR being a different approach in relationships. Relationships in a TMIR system include equivalence, hierarchical, and two types of associative relationships (both associative relationships between terms belonging to the same hierarchy and associative relationships between terms belonging to different hierarchies). Relationships in a CIR system include various relationships by clustering.

Peter Hancock		Type(s): Engineer
Untyped Names (1)	Internal Occurrences (3)	
<ul style="list-style-type: none"> • Peter Hancock 	<ul style="list-style-type: none"> • Description <ul style="list-style-type: none"> ◦ Peter works for Computer Associates. He is specialized in Intrusion Detection System. He hold CISSP. CISSP Certification was designed to recognize competency in the practice of Information Security. Certification can enhance a professional's career and affirm their level of Information Security mastery and competence. • Email <ul style="list-style-type: none"> ◦ peterh@ca.com • Telephone <ul style="list-style-type: none"> ◦ 1 415 423-1456 	
Associations (2)	External Occurrences (1)	
<ul style="list-style-type: none"> • Maintains <ul style="list-style-type: none"> ◦ Intrusion Detection System • works for <ul style="list-style-type: none"> ◦ Computer Associates 	<ul style="list-style-type: none"> • Homepage <ul style="list-style-type: none"> ◦ http://www.ca.com/peter/ 	

Fig. 3. TMIR

In other words, various associative relationships by clustering exist in CIR system. Based on ontology modeling, a TMIR and a CIR was implemented. In order to implement and navigate ontology, an ontology language and browser were required. Topic maps were used to implement ontology, and “Omnigator” as a topic maps browser was used to demonstrate what we developed for an ontology-driven information retrieval system and a clustering-based information retrieval system. Omnigator is developed by Ontopia and is a free topic map browser that allows users to navigate, test, and debug topic maps (Ontopia, 2005). The name comes from a contraction of “omnivorous navigator.” Omnigator includes a graphic visualization component based on the Vizigator and uses syntax called linear topic map notation (LTM) to build topic maps. Omnigator is based on open standard technologies, in particular XML topic maps

(XTM) and ISO 13250 (Ontopia, 2005). The TMIR and CIR interfaces were designed to be identical and to contain the same domains (Figure 3 and Figure 4).

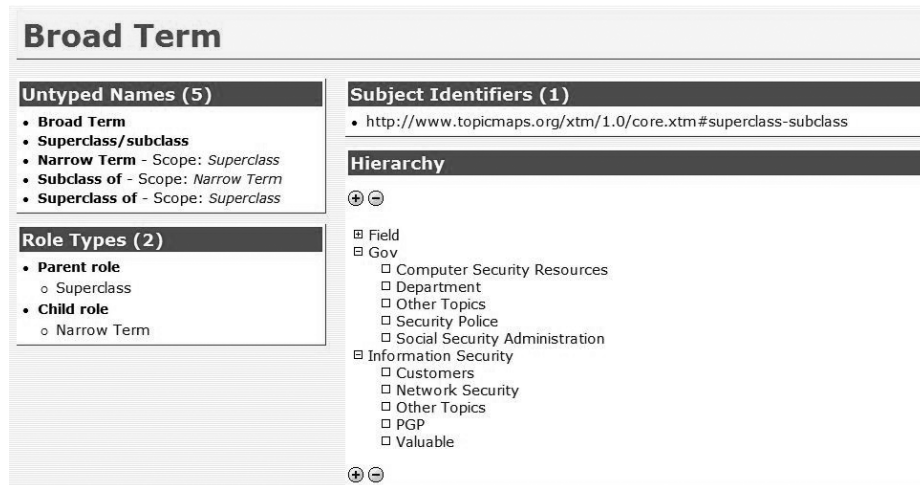


Fig. 4. CIR

6 Research Design

The primary method to address the research questions regarding both systems will be user performance testing. To evaluate the user performance, a comparative experiment will be conducted in which the performance of a TMIR system will be compared alongside a CIR system. A questionnaire about demographic and computer and search engine experiences will be formulated to screen user background. Seven queries will be formulated and distributed to twenty participants to guide their searches using both systems.

6.1 Experiment Participants

Security domain is likely to be used by network or security professionals who deal with various network or security related domains to provide services to clients. Library and Information studies undergraduate students are individuals who may be expected to pursue careers as network or security professionals in the future. Twenty participants (ten for each group) were recruited from the

students who registered for undergraduate courses from multiple higher education institutions for an experimental test.

6.2 Variables

For this study, there was one independent variable: the system (TMIR and CIR). This study included four dependent variables: recall, precision, search time, and search steps. Recall is defined as the percentage of number of relevant documents in relation to the number of relevant documents in the system. Precision is the percentage of relevant documents in relation to the number of documents retrieved. Search time is defined as the period of time devoted to looking for information for the purpose of locating relevant information in response to a request. Search step is defined as the steps to looking for information for the purpose of locating relevant information in response to a request.

6.3 Procedure

Participants conducted searches in a classroom where computers were available. The computers had identical operating systems and browsers. Participants were randomly assigned to either the experimental or the control group. The experimental group was asked to use a TMIR system to search while the control group was asked to use a CIR system. Each group's participants were given the same list of queries and were asked to perform the searches. The tasks included answering queries from the security domain.

The experiment was comprised of four sessions as follows:

1. Pre survey: A questionnaire session about demographic and computer and search engine experiences.
2. A training session including an introduction and a short practice.
3. A test session.
4. Post survey: A questionnaire session about the ease of use, satisfaction and comments

6.4 Data Collection

Two methodologies were used in this study to collect data: Questionnaires and screen recordings. Test sessions are only recorded to analyze recall, precision, search time, and search steps.

6.5 Search Tasks

Search tasks were developed from ontology-modeling. Seven search tasks were formulated and distributed to participants to guide their searches using both systems, as shown in Table 2. These tasks were categorized based on the relationship complexity. The complexity was based on the numbers of concepts, hierarchies, and the degree of relationships between concepts (Byström & Järvelin, 1995). Task categories and task assigned are as follows:

Table 2. Search Tasks

Task #	Degree of Relationships	Task
1	Simple Task	List all the security software
2	Complex Task	Name all the Security engineer who works for Cisco
3	Complex Task	Find Vendors providing security training service
4	Association and Cross Reference Related Task	List all the security hardware supported by IBM Consultants
5	Association and Cross Reference Related Task	List all software using RSA cryptography and find engineer who specializes in these software packages.
6	Association and Cross Reference Related Task	Find security system engineer(s) who specializes in firewall and their supervisor and sale representatives
7	Association and Cross Reference Related Task	Assume that you organization is interested in security training. Who will be the right people to contact? Please provide their e-mail address

To evaluate the TMIR and CIR systems, a comparative pilot study was conducted in which the performance of an Topic maps-based IR system was evaluated alongside a clustering-based IR system in order to determine and then compare

their respective recall, precision, search time, and search steps. The experimental and control groups were given the same tasks and searched for answers using two different information retrieval systems.

7 Results and Discussion

There was a significant difference in recall between the two groups. The estimate value shows the recall on TMIR was higher than CIR. The estimate value also has shown that the search time in the experimental group was less than in the control group. The three research questions were answered with the following conclusions: there were significant differences between the two groups and in terms of recall, precision, search time, and search steps. Overall, recall was higher when performing simple task than when performing complex tasks. The experimental group showed higher recall than the control group. Performing complex-tasks took more search time than performing simple tasks across the two groups. The control group took more total search time than the experimental group.

8 Conclusion

This study illustrates that the positive influences of a Topic map-based ontology IR system are improved recall/precision, shorter search time and search steps for given search tasks than the clustering-based IR system. This study shows that TMIR system resulted in better recall/precision and shorter search times/steps than CIR system. The results of this study attest to the potential of Topic Maps-based ontology to improve information retrieval system performance through better support for associative relationships between terms belonging to different hierarchies by providing explicit relationships among resources.

References

- Biren Shah, Raghavan, V., Dhatri, P., & Zhao, X. (2006). A cluster-based approach for efficient content-based image retrieval using a similarity-preserving space transformation method. *Journal of the American Society for Information Science and Technology*, 57(12), 1694-1707.
- Busygin, S., Prokopyev, O., & Pardalos, P. M. (2008). Biclustering in data mining. *Computers & Operations Research*, 35(9), 2964-2987.
- Byström, K., & Järvelin, K. (1995). Task complexity affects information seeking and use. *Information Processing and Management*, 31(2), 191 - 213
- Dunlavy, D. M., O'Leary, D. P., Conroy, J. M., & Schlesinger, J. D. (2007). QCS: A system for querying, clustering and summarizing documents. *Information Processing & Management*, 43(6), 1588-1605.
- E.K.F. Dang, Luk, R. W. P., Ho, K. S., Chan, S. C. F., & Lee, D. L. (2008). A new measure of clustering effectiveness: Algorithms and experimental studies. *Journal of the American Society for Information Science and Technology*, 59(3), 390-406.
- Garshol, L. M. (2002). What Are Topic Maps Retrieved February 12, 2006, from <http://www.xml.com/pub/a/2002/09/11/topicmaps.html>
- Hu, G., Zhou, S., Guan, J., & Hu, X. (2008). Towards effective document clustering: A constrained K-means based approach. *Information Processing & Management*, 44(4), 1397-1409.
- ISC (2008). CISSP® - Certified Information Systems Security Professional Retrieved June 2, 2008, from <https://www.isc2.org/cgi-bin/content.cgi?category=97>
- Kang, I.-S., Na, S.-H., Kim, J., & Lee, J.-H. (2007). Cluster-based patent retrieval. *Information Processing & Management*, 43(5), 1173-1182.
- Lin, Y., Li, W., Chen, K., & Liu, Y. (2007). A Document Clustering and Ranking System for Exploring MEDLINE Citations. *Journal of the American Medical Informatics Association*, 14(5), 651-661.
- Liz Price, & Thelwall, M. (2005). The clustering power of low frequency words in academic Webs. *Journal of the American Society for Information Science and Technology*, 56(8), 883-888.
- Na, S.-H., Kang, I.-S., & Lee, J.-H. (2007). Adaptive document clustering based on query-based similarity. *Information Processing & Management*, 43(4), 887-901.
- Niall Rooney, Patterson, D., Galushka, M., Dobrynin, V., & Smirnova, E. (2008). An investigation into the stability of contextual document clustering. *Journal of the American Society for Information Science and Technology*, 59(2), 256-266.
- Nosovski, G. V., Liu, D., & Sourina, O. (2008). Automatic clustering and boundary detection algorithm based on adaptive influence function. *Pattern Recognition*, 41(9), 2757-2776.
- Ontopia (2005). The Ontopia Omnigator: User's Guide. Retrieved from <http://www.ontopia.net/download/index.html>

- Oscar Loureiro, & Siegelmann, H. (2005). Introducing an active cluster-based information retrieval paradigm. *Journal of the American Society for Information Science and Technology*, 56(10), 1024-1030.
- Ronald N. Kostoff, & Block, J. A. (2005). Factor matrix text filtering and clustering. *Journal of the American Society for Information Science and Technology*, 56(9), 946-968.
- Sherry Koshman, Spink, A., & Jansen, B. J. (2006). Web searching on the Vivisimo search engine. *Journal of the American Society for Information Science and Technology*, 57(14), 1875-1887.
- VicencTorra, Lanau, S., & Miyamoto, S. (2006). Image clustering for the exploration of video sequences. *Journal of the American Society for Information Science and Technology*, 57(4), 577-584.