Metadata and Semantics: A Case Study on Semantic Searching in Web System

Marut Buranarach[1] and Michael B. Spring[2]
1 National Electronics and Computer Technology Center (NECTEC)
Thailand Science Park, Phatumthani 12120, Thailand
marut.buranarach@nectec.or.th,
WWW home page: http://www.nectec.or.th/
2 Department of Information Science and Telecommunications,
University of Pittsburgh
Pittsburgh, PA 15260, USA
spring@imap.pitt.edu,
WWW home page: http://www.sis.pitt.edu/~spring/

Abstract. Metadata is used in Web information systems to improve search results. Searching for resources based on metadata usually relies on keyword matching, i.e. user's query terms must match with metadata terms of the relevant resources. The Semantic Web framework incorporates ontology with metadata to retrieve resources. This paper describes a case study of semantically enhanced searches in a Web information system. The study combines ontologies with metadata to augment searches in a Web resource collection. Ontologies representing semantic associations between the US presidents were created and allow some semantic queries for the domain. For example, a semantic query for resources about "the US president who is the successor of Bill Clinton" can be formed instead of the query term "George W. Bush". The semantic queries were formed using description logic expressions. Evaluation of the results demonstrated good precision with moderate recall. Factors contributing to and detracting from retrieval performance were identified and addressed. The results highlight both the potential and issues in combining ontologies with metadata for enhancing search in next-generation Web information systems.

Please use the following format when citing this chapter:

Buranarach, M., Spring, M., B., 2006, in International Federation for Information Processing, Volume 205, Research and Practical Issues of Enterprise Information Systems, eds. Tjoa, A.M., Xu, L., Chaudhry, S., (Boston:Springer), pp.507-517.

1 Introduction

Metadata is a surrogate for a resource represented in terms of attribute name-value pairs. These attributes provide additional information about the resource such as title, author, subject, etc. Metadata allows the retrieval of resources by matching terms in user queries with terms in the specified metadata attributes. For example, in retrieving the resources authored by John Smith, the query term: *author="John Smith"* may be used. Some applications of metadata in Web information systems range from searching multimedia resources to searching product catalogs.

One of the limitations of metadata search is that the keywords used in user queries must match with the keywords used in the metadata. For example, in order to find all the resources whose subject is about the US president "George W. Bush", the query term: *subject="George W. Bush"* must be used. Metadata by itself does not allow semantic queries, i.e. queries that rely not only on keyword matching but on underlying semantics. For example, the resources found for the above query are also resources for queries about "the current US president", "the US president who is the successor of Bill Clinton" or "the US president who is a son of another president". These queries share similar meaning to the first query but cannot be used in either full-text or metadata based systems.

Semantic search is a new approach for search enabled by the Semantic Web framework [1]. Semantic search offers an enhancement to keyword-based search in that the words used in queries do not need to match the words used in describing the resources. In particular, it allows for retrieval that incorporates the underlying semantics of terms. Ontologies are used in most semantic search systems. The potential of semantic search has been demonstrated in some search systems [2-4].

This paper presents a case study on the use of semantic search in a Web information system. It demonstrates the combined use of ontology and metadata in enabling semantic search in a Web resource collection. Unlike prior work, it quantifies the retrieval effectiveness obtained based on expert analysis of the corpus and the retrieval results. Metadata were extracted from some book titles in the collection of Amazon.com. Ontologies on the subject topic of the US presidents were created and integrated to augment metadata search on the subject. The ontologies represented semantic associations between the US presidents, i.e. their chronological orders and biological relationships. A semantic search system was built over a description logic system. Nine queries for books about the US presidents' biography were formed based on the semantic associations. Evaluation of the results demonstrated the value of semantics in Web searches. Some factors that impacted the retrieval performance were identified and addressed.

2 Background

Description logic (DL) is often used for the logic layer of the Semantic Web. Specifically, the Web Ontology Language (OWL) standard contains the OWL DL

sub-language providing expressiveness that supports inference by description logic [5]. Description logic consists of three basic types: *Concept*, *Role* and *Individual*. A concept can be *primitive* or *defined*. A concept is a *defined* concept, if it can be described in term of previously known concepts; otherwise it is a *primitive* concept. A role is property of concept. An individual is similar to concept but can only be used to describe at most one individual. Below is some background on DL theory summarized based on [6-8].

The semantics of DL is usually given using the notion of interpretation. The interpretation $I = (\Delta^I, {}^I)$ consists of a non-empty set (Δ^I) and an interpretation function $({}^I)$. The interpretation function could be applied to a concept, i.e. $C^I = I(C)$, which maps a concept into a subset of Δ^I. The interpretation could be applied to a role, i.e. $R^I = I(R)$, which maps a role into a subset of the cartesian product of Δ^I, i.e. $(\Delta^I \times \Delta^I)$. The interpretation function could be applied to an individual, i.e. $O^I = I(O)$, which maps an individual name into a member of Δ^I.

$\mathcal{SHIQ}$ is an expressive description logic, whose expressiveness also includes role transitivity ($\mathcal{R}+$), hierarchy of role ($\mathcal{H}$) and inverse role ($\mathcal{I}$). $\mathcal{SHIQ}$ concept expressions can be constructed using the combination of the following constructors: $\neg C$, $(C \sqcap D)$, $(C \sqcup D)$, $(\exists R.C)$, $(\forall R.C)$, $(\leq n\ R.C)$ and $(\geq n\ R.C)$, where C, D are concepts, R is a role, and n is an integer. A role is a transitive role if it satisfies the following condition: if $(x, y) \in R^I$ and $(y, z) \in R^I$, then $(x, z) \in R^I$. The syntax and semantics of $\mathcal{SHIQ}$ concepts and roles are provided in Fig. 1.

Concepts		
Syntax	**Description**	**Semantics**
A	Concept name	$A^I \subseteq \Delta^I$
$\neg C$	Negation	$\Delta^I \setminus C^I$
$C \sqcap D$	Conjunction	$C^I \cap D^I$
$C \sqcup D$	Disjunction	$C^I \cup D^I$
$\exists R.C$	Existential quantification	$\{x \mid \exists\, y\ (x, y) \in R^I \wedge y \in C^I\}$
$\forall R.C$	Universal quantification	$\{x \mid \forall\, y\ (x, y) \in R^I \Rightarrow y \in C^I\}$
$\leq n\ R.C$	Qualified number restriction	$\{x \mid \#\{\, y \mid (x, y) \in R^I \wedge y \in C^I\} \leq n\}$
$\geq n\ R.C$		$\{x \mid \#\{\, y \mid (x, y) \in R^I \wedge y \in C^I\} \geq n\}$

Roles		
Syntax	**Description**	**Semantics**
R	Role name	$R^I \subseteq \Delta^I \times \Delta^I$
R^{-1}	Inverse role	$\{(x, y) \in \Delta^I \times \Delta^I \mid (y, x) \in R^I\}$

Fig. 1. Syntax and semantics of $\mathcal{SHIQ}$ concepts and roles

TBox Statements	
Syntax	**Satisfied if**
$C \doteq D$	$C^I = D^I$
$C \sqsubseteq D$	$C^I \subseteq D^I$
$R \sqsubseteq S$	$R^I \subseteq S^I$

ABox Statements	
Syntax	**Satisfied if**
C(a)	$a^I \in C^I$
R(a, b)	$(a^I, b^I) \in R^I$

Fig. 2. Syntax and semantics of *TBox* and *ABox* statements

A $\mathcal{SHIQ}$ knowledge base ***K*** consists of two kinds of statements: *terminological* and *assertional*. The set of the first kind of statements constitutes the ***TBox***. The set of the second kind constitutes the ***ABox***. The ***TBox*** contains the statements describing concepts and roles. The ***ABox*** contains the statements describing individuals. ***TBox*** and ***ABox*** statements are in the forms shown in Fig. 2.

The first form of ***TBox*** statements ($C \doteq D$) indicates equivalence between two concepts. The second form ($C \sqsubseteq D$) indicates a subsumption relationship between two concepts. In particular, concept C is subsumed by concept D if every individual that is a member of concept C is also a member of concept D. The third form ($R \sqsubseteq S$) indicates a subsumption relationship between two roles. The first form of ***ABox*** statements (C(a)) indicates that individual *a* is a member of concept *C*. The second form (R(a, b)) indicates that two individuals: *a* and *b* are related by role *R*.

3 Implementation

3.1 System Architecture

A conceptual architecture for implementing semantic search for a Web information system is shown in Fig. 3. The four major components include Web Resource Collection, Semantics Data Source, Semantic Search System and User Queries. Web resource collection is where resources and metadata reside. Semantics data source provides ontologies. In this study, the resource collection was autonomous. The ontologies were created by the study and independent of the resource collection. Semantic search system was built over a description logic system. It acquired metadata from the resource collection and ontologies from the semantics data source. The acquired data were pre-processed, e.g. parsed and reformatted, before they were interpreted by the description logic system. The user submits a query as a description logic expression to the system. The system returns a list of resources whose metadata semantics match with query semantics. The results were not ranked.

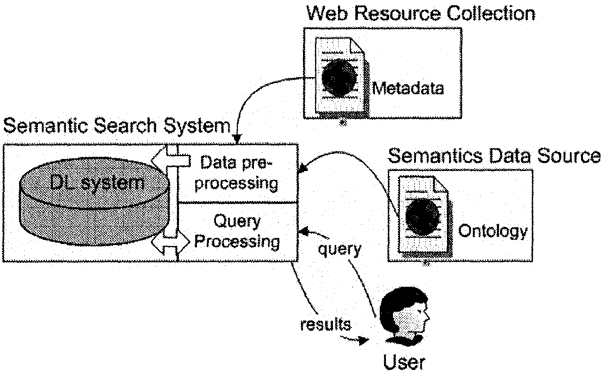

Fig. 3. Conceptual architecture of semantic search for a Web information system

3.2 Semantic Search System

The implementation of semantic search system utilizes the RACER (Renamed ABox and Concept Expression Reasoner) system [7]. The RACER system is a knowledge representation system that has a support for the description logic $\mathcal{ALCQHI}_{R+}$ or $\mathcal{SHIQ}$. All the standard inference services for ***TBox*** and ***ABox*** are supported by RACER. The implemented system utilizes the RACER system version 1.7.6 running on a computer with Windows 2000 operating system, Pentium-733MHz and 1GB RAM.

3.3 Web Resource Collection

The study used the book collection of the Amazon.com website as the Web resource collection. Only the book titles in the subject category: "*Biographies&Memoirs/ Leaders&NotablePeople/ Presidents&HeadsOfState*"[3] were used by the study. The category is the most likely category containing biography books about the US presidents, which was the subject topic for the study. Metadata for 927 book titles in the subject category was acquired via the Amazon.com Web service interface[4] on March 21, 2004. The obtained data was cached to insure the consistency of the data used across different queries.

Metadata for each book title were pre-processed and represented to the DL system in the following format:

[3] http://www.amazon.com/exec/obidos/tg/browse/-/2418
[4] http://www.amazon.com/webservices/

has_subject (i1234567890, george_washington)

Where *i1234567890* is the book's ISBN plus the 'i' prefix, *george_washington* is a keyword assigned to the metadata attribute "subject" for this book title, where every whitespace character is replaced by '_' character. Only the metadata attribute 'subject' was processed for each book tile. Other metadata attributes such as author, price, etc. were ignored. The role *has_subject* was created as a primitive role in the DL system.

3.4 Ontologies

Ontologies in the subject domain of the US presidents were created. The ontologies represented semantic associations between the US presidents. In particular, the relationships between each US president, i.e. chronological order and biological relationships, were expressed. The ontologies were represented to the DL system in the following formats.

is-next-predecessor-of (bill_clinton, george_w_bush)
C_george_w_bush (george_w_bush)
C_george_w_bush $\sqsubseteq$ C_us_president

In the first statement, President Bill Clinton was modeled as the predecessor of President George W. Bush. The second and third statements represent the subsumption relationship between the concept of President George W. Bush and the concept of US president. The relationships between the US presidents defined in the ontologies are shown in Fig. 4.

In order to permit inferences, roles were modeled using inverse, transitivity and role hierarchy. For example, *is-child-of* was modeled as inverse of *is-father-of*, *is-predecessor-of* was modeled as a transitive role and *is-cousin-of* was modeled as a role subsumed by *is-relative-of*, etc. The roles defined are shown in Fig. 5.

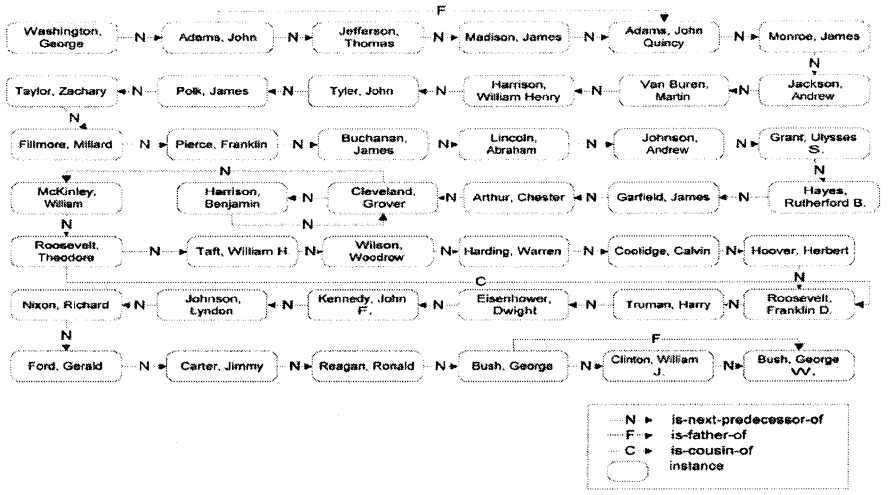

Fig. 4. Relationships between the US presidents defined in the ontologies

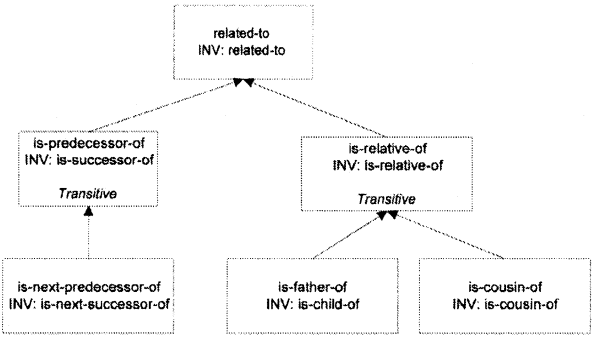

Fig. 5. Role hierarchy and properties defined in the ontologies

3.5 Queries

Nine queries in the topic area were created. The queries were expressed in terms of semantic relationships between the US presidents instead of names. Queries were defined on two aspects: chronological order and biological relationships. Four of the queries were expressed in terms of the US presidents' chronological order (Q2-Q5). Four of the queries were expressed in terms of the US presidents' biological

relationships (Q6-Q9). List of the queries and their mapping into DL expressions is shown in Table 1.

Table 1. List of the queries and their mappings into DL expressions

Queries	DL expressions
Q1: Books on biography of the US presidents	∃ has_subject.C_us_president
Q2: Books on biography of the first US president	∃ has_subject.(C_us_president ⊓ ≤ 0 is-next-successor-of)
Q3: Books on biography of the US presidents after President John F. Kennedy	∃ has_subject.(C_us_president ⊓ ∃ is-successor-of.C_john_f_kennedy)
Q4: Books on biography of the US presidents between President John F. Kennedy and President Ronald Reagan	∃ has_subject.(C_us_president ⊓ ∃ is-successor-of.C_john_f_kennedy ⊓ ∃ is-predecessor-of.C_ronald_reagan)
Q5: Books on biography of the US presidents before President Thomas Jefferson	∃ has_subject.(C_us_president ⊓ ∃ is-predecessor-of.C_thomas_jefferson)
Q6: Books on biography of the US presidents who are fathers of other US presidents	∃ has_subject.(C_us_president ⊓ ∃ is-father-of.C_us_president)
Q7: Books on biography of the US presidents who are sons of other US presidents	∃ has_subject.(C_us_president ⊓ ∃ is-child-of.C_us_president)
Q8: Books on biography of the US presidents who are cousins of other US presidents	∃ has_subject.(C_us_president ⊓ ∃ is-cousin-of.C_us_president)
Q9: Books on biography of the US presidents who are relatives of other US presidents	∃ has_subject.(C_us_president ⊓ ∃ is-relative-of.C_us_president)

4 Evaluation

The retrieval performance of the semantic search system against the queries was assessed in terms of precision and recall. The relevancy of the retrieved resources to the queries was assessed by a panel of three judges. The judges were the graduates from the Master of Library and Information Science program at the University of Pittsburgh. A resource was judged as relevant to a query if at least two judges marked it as relevant. Precision of the results for each query is shown in Fig. 6. The number of relevant resources retrieved per the number of resources retrieved for each query is also displayed in the graph.

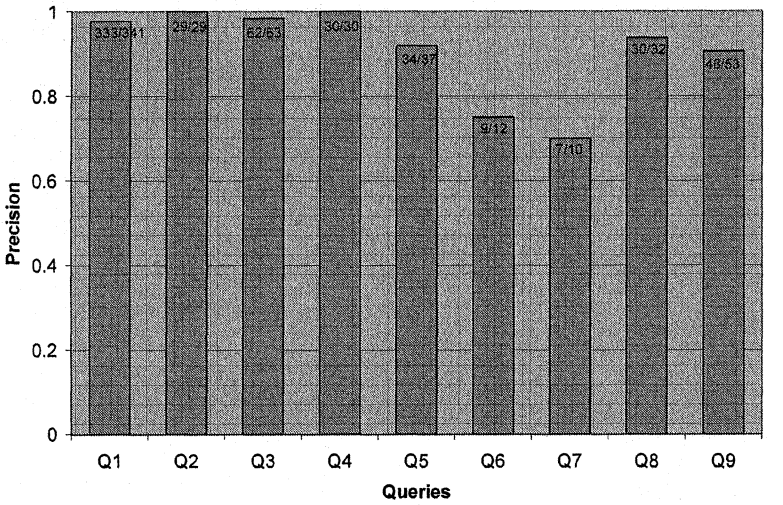

Fig. 6. Precision of the results

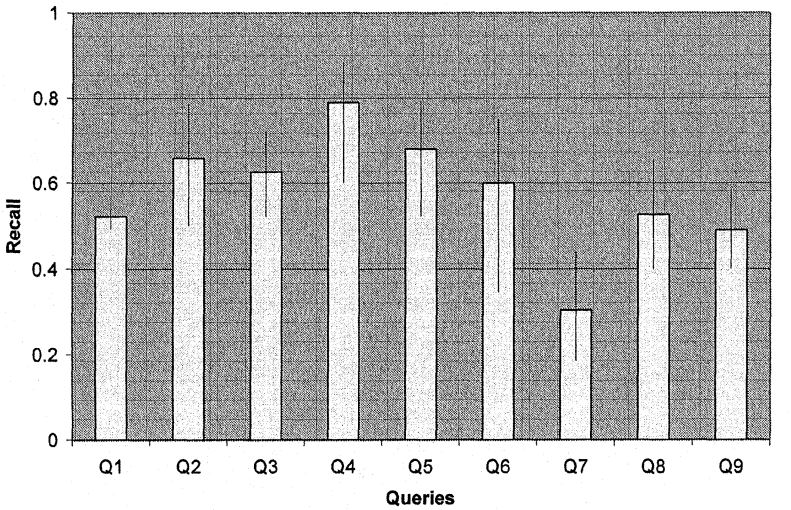

Fig. 7. Estimated recall of the results

In measuring recall, the number of relevant resources for each query must be known. However, obtaining such a number requires exhaustive examination of all

the resources in the subject category for each query. In order to provide some indication of recall while maintaining manageable number of resources reviewed by the judges, the study measured estimated recall instead of actual recall. The number of relevant resources for each query was estimated based on the assessment of 500 resources sampled randomly from the resources in the subject category[5]. The formula in measuring recall is provided as:

$$\text{Recall} = \frac{\text{Total relevant resources retrieved}}{\text{Total relevant resources retrieved} + \text{Total relevant resources non - retrieved}}$$

The number of relevant resources retrieved was the same number obtained when measuring precision. The total relevant resources non-retrieved is estimated as $p_{nr}N$, where p_{nr} is the proportion of relevant resources non-retrieved found in the sample of 500 and N is total number of resources (= 927). The estimation ranges from $L_{nr}N$ to $U_{nr}N$, where L_{nr} and U_{nr} are the lower and upper limits of the 95% confidence interval for p_{nr} and are obtained using the formulas defined in [9]. The estimated recall and the estimation range for each query is shown in Fig. 7.

The results show good precision (0.7-1.0, average = 0.91). Some degradation in precision was due to some questionable classification of resources and resources with less degree of relevancy. In particular, some resources located in the subject category were questionably not biography books about the US presidents. For example, the subject category includes some books written about letters from the US presidents, which were not considered biography books about the US presidents. Other possible misclassifications include books about first ladies or other family members included in the subject category of the US presidents. The most degraded precision was found in Q6 (= 0.75) and Q7 (= 0.7). These were due to books focusing on the story of President John Adams' entire family, not only the members that were US presidents. These resources were assessed as irrelevant to the queries due to their generality.

The results show overall moderate recall (0.3-0.79, average e= 0.58). Degradation of recall was due to the use of general subject terms in resource metadata. In particular, many books in the subject category were assigned the general subject term "Presidents and Heads of State" instead of particular names of the presidents that the books are about. For example, many books about President George W. Bush were not retrieved because they were assigned only the general subject term. This resulted in the most degraded recall in Q7 (= 0.3). The lack of specific information in the subject terms was the only cause found for the degraded recall.

[5] The evaluation scheme was developed as a part of a larger study involving more queries and resources. The reduction in effort was not substantial in the subset reported in this paper.

5 Conclusions

This paper presents a case study on semantic searching in a Web information system. A semantic approach in querying resources about the US presidents was demonstrated. Ontologies were combined with metadata search to allow for expressive semantic-based queries. Evaluation of the retrieval system over the defined queries showed promising retrieval performance, particularly in terms of precision. Retrieval performance was impacted by some errors and omissions in the metadata of the resource collection. Improving the metadata quality could result in improved retrieval performance by the system. In particular, misclassification of resources in the collection should be minimized to improve precision. Resource metadata should be provided as specifically as possible to improve recall.

The results highlight both the potential of semantic search in Web information system and some factors that can impact its retrieval performance. Other related issues such as user interfaces and result ranking are beyond the scope of this paper and should be further investigated.

References

1. W3C Semantic Web (December 8, 2005); http://www.w3.org/2001/sw/.
2. C. Rocha, D. Schwabe, and M.P. de Aragao, A Hybrid Approach for Searching in the Semantic Web, In the Proceedings of the World Wide Web Conference, pp. 374-383 (2004).
3. A. Sheth and C. Ramakrishnan, Semantic (Web) Technology in Action: Ontology Driven Information Systems for Search, Integration and Analysis, *IEEE Data Engineering Bulletin* **26**(4), 40-48 (2003).
4. R. Guha, R. McCool, and E. Miller, Semantic Search, In the Proceedings of the World Wide Web Conference, pp. 700-709 (2003).
5. D.L. McGuinness and F. van Harmelen, OWL Web Ontology Language Overview (February 10, 2004); http://www.w3.org/TR/owl-features/.
6. I. Horrocks and S. Tessaris, Querying the Semantic Web: a Formal Approach, In the Proceedings of the 2002 International Semantic Web Conference (ISWC 2002), edited by I. Horrocks and J. Hendler (Springer-Verlag, 2002).
7. V. Haarslev and R. Möller, Description of the RACER System and Its Applications, In the Proceedings of the 2001 International Description Logics Workshop (DL-2001), pp. 132-141 (2001).
8. S. Tessaris, Questions and Answers: Reasoning and Querying in Description Logic, Ph.D Dissertation, University of Manchester (2001).
9. R.G. Newcombe, Two-Sided Confidence Intervals for the Single Proportion: Comparison of Seven Methods, *Statistics in Medicine* **17**, 857-872 (1998).