

OntoKhoj: A Semantic Web Portal for Ontology Searching, Ranking and Classification

Chintan Patel, Kaustubh Supekar, Yugyung Lee, E.K. Park
School of Computing and Engineering
University of Missouri-Kansas City
{copdk4, kss2r6, leeyu, ekpark@umkc.edu}

ABSTRACT

The goal of the next generation Web is to build virtual communities, wherein software agents and people can work in cooperation by sharing knowledge. To achieve this goal, the emerging Semantic Web community has proposed ontologies to express knowledge in a machine understandable way. The process of building and maintaining ontologies, which is known as Ontology Engineering, presents unique challenges. These challenges are related to lack of trustworthy and authoritative knowledge sources and absence of a centralized repository to locate ontologies to be reused. In this paper, we propose a Semantic Web portal, called OntoKhoj that is designed to simplify the Ontology Engineering process. The methodology in developing OntoKhoj is based on algorithms used for searching, aggregating, ranking and classifying ontologies in Semantic Web. The proposed OntoKhoj would 1) allow agents and ontology engineers to retrieve trustworthy, authoritative knowledge, and 2) expedite the process of ontology engineering through extensive reuse of ontologies. We have implemented the OntoKhoj portal and further validated our system on the real ontological data in the Semantic Web.

Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]: Query formulation, Retrieval models, Search process

General Terms

Management, Design, Standardization

Keywords

Semantic Web, Searching, Ranking, Classification

1. INTRODUCTION

The Semantic Web is an emerging field, with the aim of building infrastructure, wherein software agents and peo-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

WIDM'03, November 7–8, 2003, New Orleans, Louisiana, USA.
Copyright 2003 ACM 1-58113-725-7/03/0011 ...\$5.00.

ple can work in cooperation by sharing knowledge [1]. This requires incorporating machine understandable information into the Web designed solely for human consumption. With the support of a new set of solutions developed by the Semantic Web community to meet these requirements, more Web content represented in ontologies would be accessible to machines. The process of building and maintaining ontologies in the Semantic Web, which is known as Ontology Engineering, presents unique challenges.

In this paper, we try to tackle the two major challenges: 1) Searching ontologies and 2) Trusting Information over the Semantic Web. First, the Semantic Web is facing the same problems encountered by WWW during its nascent stages, namely searching relevant information over Web. In the Semantic Web, ontologies represent the knowledge to be shared by formally defining concepts and relations of entities occurring in a domain or universe of discourse.

Another issue is the amount of trust we place on the information present on Semantic Web. In an independent environment such as Web, where there are no restrictions on the information being published, it becomes the responsibility on the part of the consumer to accurately judge the quality and validity of the information provider. Semantic Web has been envisioned to allow machines to work along with or on behalf of human beings. This implies machines are responsible, to a certain degree, for discerning the trustworthiness of information source. Currently, there are no appropriate solutions to characterize validity and quality of ontologies.

We are highly motivated by the fact that having a dynamic and trustworthy ontology information source is extremely important in advancement and growth of Semantic Web. To visualize the solutions to aforementioned problems we draw an analogy from current Web: Ontologies in Semantic Web are akin to Web pages connected to each other using hyperlinks aka relationships (rdf:about, rdfs:subClass). In current Web, information searching and indexing is performed by specialized search engines (e.g. Google.com, Altavista.com) using proprietary algorithms to crawl and rank the Web pages and subsequently allowing users to perform simple query, keyword based searches. Along similar lines we have developed a Semantic Web portal, OntoKhoj, that would provide services related to searching, ranking, aggregating and classifying ontologies crawled from the Semantic Web, thereby providing Knowledge Engineers and software agents, a source for authoritative, trustworthy ontologies.

2. RELATED WORK

Growth of Semantic Web has led to massive growth in

the use and development of ontology. The central idea of Ontology Engineering in Semantic Web is extensive reuse of existing ontologies. Currently, Semantic Web doesn't have any infrastructure that allows Knowledge Engineers to search and peruse relevant domain ontologies. Lack of central index of ontologies aggravates the problem. Recently, several tools for Ontology Engineering have been developed: Protégé-2000¹, OntoEdit², OilEd³. However, these tools do not provide any facilities for Knowledge Engineers to share or collaborate and reuse their work.

Ranking based on citations has been a major area of research [3, 2]. Google⁴ PageRank is one of finest examples that shows the success of the citation algorithms [6] in Web environment. OntoKhoj extends the functionality of PageRank to handle the critical issues that arise while considering Web of ontologies rather than simple Web of HTML (Hypertext Markup Language) pages.

Another significant initiative in this area is Ontolingua [5] that gives a distributed collaborative environment to browse, create, edit, modify, and use ontologies, but it requires the user to register and then publish ontologies. OntoKhoj uses Web and Semantic Web crawling techniques to retrieve ontologies thereby preserving spirit of openness and independence of publishing information over Web.

3. THE ONTOKHOJ MODEL

3.1 Ontology Crawling & Classification

Performing *knowledge crawling*, which would be quite different from Web page crawling, requires consideration of specific features in the underlying knowledge representation framework. Semantic Web is based on Resource Description Format (RDF) model⁵, wherein each RDF entity has an associated URI⁶ that allows citation and reusability, thereby accelerating the proliferation of knowledge. The Semantic Web also allows ontologies to be distributed, i.e. several RDF segments belong to same logical URI but physically present at different locations. After crawling, we aggregate those segments into a single ontology, depending upon the URI of the concept.

Domain Ontologies are to be specific to a particular domain and henceforth capturing most of the common terminologies for the given domain. For Ontology classification, Ontologies can be assumed as plain texts that contain terms describing concepts and relationships. Thus, we applied traditional classification algorithms to Ontology classification. In OntoKhoj, the classification model was trained by the data derived from DMOZ⁷, which contains a large number of manually classified datasets. Then, it is determined, based on the model, whether the ontology belongs to a particular topic. The classified ontologies are stored into the corresponding directory and can be graphically presented to users, and also can be traversed or retrieved by agents. We will show that the experimental results confirm the effectiveness of our classification approach in Section 5.

¹<http://protege.stanford.edu/>

²<http://www.ontoknowledge.org/tools/ontoedit.shtml>

³<http://oiled.man.ac.uk/>

⁴<http://www.google.com/>

⁵<http://www.w3.org/RDF/>

⁶<http://www.ietf.org/rfc/rfc2396.txt>

⁷<http://www.dmoz.org/>

| Priority (Weight) | Relationship | Language Specific |
|-------------------|---------------|------------------------------|
| 1 | instantiation | rdf:type |
| 2 | subClass | rdfs:subclass, daml:subClass |
| 3 | domain/range | rdfs:domain, daml:range |

Table 1: Weights of Hyperlinks

3.2 Ontology Ranking

As the Semantic Web is much more complex than traditional hyperlinked Web pages, we focus on the type of relationships, which exist in Semantic Web languages (RDF, RDFS, DAML+OIL, OWL). The various types of hyperlinking across ontologies have some inherent semantics that could be exploited to determine their importance. For example, if a concept is a subconcept of another concept in a different ontology, it indicates that some original features are maintained, while new features added to the concept. If a concept instantiates another one in a different ontology, it is an endorsement which ensures a much strong reference. Referring to a concept in another ontology as domain/range would assume the least priority. As depicted in Table 1, we prioritized such semantic relationships based on the intuitive reasoning described above.

Crawler fetches the RDF documents according to the physical links (HTML URLs, RDF URIs). However, since URI may not necessarily point to an actual physical resource, we modeled a logical layer consisting of hyperlinked RDF ontologies. Based on the logical layer, we propose an ontology ranking methodology which considers the idea of *Referencing*. For a given concept C_i , in ontology O_i , and the ontology referencing hyperlinks $Ref \in \{rdf:type, rdfs:subclass, rdfs:domain, rdfs:range\}$. Other properties (rdf:seeAlso, rdf:about) may be considered in our model. We define the following terms:

- The identity and rank of the referrer, $Ref(?C, C_i)$, where $?C$ is a undetermined concept.
- The number of citations by others, $|Ref(?C, C_i)|$
- The distance of reference from the origin C_o to the target C_d , $Dist(Ref(C_o, C_d))$ where a chain of referring exists from C_o to C_d (e.g., $Ref(C_o, C_i), \dots, Ref(C_j, C_d)$, avoiding a circular reference such as $Ref(C_o, C_d)$ and $Ref(C_d, C_o)$). Thus, the distance of a direct referencing is equal to 1. We place less weight on concepts that are further apart through the reference links.

Our work is influenced by the PageRank algorithm [6], which measures its citation importance by using maps containing minimum 518 million hyperlinks, and prioritizes the results of keyword-based searches in the Google system. Similarly, we have developed an algorithm OntoRank that assigns a rank to an ontology in Semantic Web. However, our work can be distinguished from the PageRank [6] by a number of special aspects such as considering different types of link and additional constraints like distances. Hence apart from considering the rank of the referrer, we also take into consideration the weight of the type of references.

We give a formal treatment to the aforementioned methodology. Let O be the ontology whose rank is to be determined. Let α be the number of ontologies referring O , each of the referring ontologies can have more than one referrals. Let β_i be the number of referrals from ontology O_i to O . Let Ω_i

be the total number of outgoing referrals from ontology O_i . Let T be the weight of the reference, N be the normalization factor. The OntoRank, $OR(O)$ is defined as follows:

$$OR(O) = N * \sum_{i=1}^{\alpha} (1/\Omega_i) * \sum_{j=1}^{\beta_i} OR(O_i) * T_j \quad (1)$$

We believe that the simplicity of the proposed algorithm accounts for its scalability and stability. We forego many other issues related to circular references, dangling links for the sake of brevity.

3.3 Ontology Searching

With the growth of ontologies over the Web, one of the challenging tasks is to search for the desired ontologies. To the best of our knowledge, there are no engines that can search ontologies on behalf of the users of Semantic Web (i.e., humans and machines). Studies on Web search interface show that users prefer simple keyword-based interfaces over complex query-based interfaces. However, the results from a simple keyword-based query may be too insufficient to determine the right query context, leading to poor precision. In order to meet the requirements, OntoKhoj provides several different interfaces to satisfy search requirements of users at different conceptualization levels.

Context Oriented Query Interface: Our approach for constructing the Context Oriented Query Interface is based on three dictionary entries: senses, synonyms, and hypernyms. Our approach relies on WordNet that is a lexical ontology developed by Miller et al. [4]. The OntoKhoj portal has an interface to the WordNet lexical reference system, which is responsible for the retrieval and display of the dictionary entities. The operations of Context Oriented Query Interface are summarized as follows:

- The search interface allows the user to disambiguate senses by selecting a correct sense from the displayed listing of various senses associated with the keyword. A concept may have different meanings in different contexts (e.g. concept *date* could be referred to as *day of the month* or *sweet edible fruit of the date palm with a single long woody seed*).
- For the selected sense of the keyword, associated synonyms and hypernyms are retrieved from the WordNet.
- If the search term is not an exact match for any of the concepts in the ontologies, the closest (synonym) matching is performed.
- If the search term is not found, then the hypernymic matching is performed. It traverses the hypernymic link upward until a term which is close to the keyword of interest is found.

OntoKhoj Machine Interface: Semantic Web is meant for agents to interpret information on Web in lieu for humans. In this spirit we need to automate the process of searching ontologies and possibly interpret information on users' behalf.

- An interface for agents to access and query the directory of classified ontologies is provided. The directory is represented in RDF ontology that allows agents to

automatically traverse and retrieve desired information.

- Advanced Logic query interfaces (e.g. RDQL, FLogic) that allows specifying search constraints, thereby providing sophisticated inferencing capabilities across ontologies.

4. THE ONTOKHOJ IMPLEMENTATION

We have implemented OntoKhoj, a Semantic Web Portal, that is designed to simplify the Ontology Engineering process. The implementation methodology is based on algorithms used for searching, crawling, classifying and ranking ontologies in Semantic Web. In current Semantic Web, multiple ontologies describing a same domain/concept appear to be quite common. Responding to the urgent needs of the Semantic Web in the current context, the implemented OntoKhoj portal 1) allows agents and ontology engineers to retrieve trustworthy, authoritative knowledge, and 2) expedites the process of Ontology Engineering through extensive reuse of ontologies. The tool is currently accessible through our website at <http://sice527.ddns.umkc.edu/ontokhoj>.

The prototype system of the OntoKhoj portal was implemented using Java on Linux platform. The four major functionalities include 1) Crawling ontologies over the Web, 2) Ranking these crawled ontologies in a local repository, 3) Classifying each of the stored ontology, 4) Ontology Searching and Visualization.

The first task of crawling ontologies was accomplished through RDF crawler, which combines advanced features of Ontobroker RDF Crawler⁸ and Jena⁹. Our crawler can retrieve ontologies represented in RDF, RDF embedded HTML and DAML+OIL format. The crawled ontologies are stored in a local repository, which is a MySQL database. As the second task, the crawled ontologies are indexed and ranked for determining authoritative ontologies.

As the third task of classification, our Java based implementation extracts the corresponding text from the DMOZ repository, wherein 460,000 Web page categories are classified and each category is associated with about 100 Web pages. Each of the ontologies stored in the OntoKhoj local repository is entered into the trained Rainbow tool¹⁰, a document classification tool; subsequent testing yields a classification of the selected ontology. The ontology is classified using four classification algorithms - Naïve Bayes, TFIDF, Probabilistic Indexing (PRIND) and K-Nearest neighbor (KNN). Finally, the OntoKhoj portal provides users with a Web-based ontology search interface and visualization tool, implemented based on GraphViz¹¹, which converts the classified ontologies into a visual representation.

5. EXPERIMENTAL RESULTS

The RDF crawler was executed for 48 hours and it yielded considerable amount of data (Table 2). Due to our limited computational resources, the execution period of RDF crawler was restricted to 48 hours. Further, as the number of publicly available ontologies is limited, we consider

⁸<http://ontobroker.semanticweb.org/rdfcrawl/>

⁹<http://www.hpl.hp.com/semweb/jena.htm>

¹⁰<http://www-2.cs.cmu.edu/mccallum/bow/>

¹¹<http://www.research.att.com/sw/tools/graphviz/>

| | |
|--------------------------------------|---------|
| Number of Web pages visited | 2018412 |
| Number of Concepts crawled | 19870 |
| Number of Relationships Discovered | 1321 |
| Total Ontologies (after Aggregation) | 418 |

Table 2: OntoKhoj statistics

| Domain | Classification | Precision | Recall |
|------------------|----------------|-----------|--------|
| University | Naïve Bayes | 1.0 | 1.0 |
| | TFIDF | 0.8 | 0.66 |
| | KNN | 0.83 | 0.83 |
| | PRIND | 0.6 | 1.0 |
| Computer Science | Naïve Bayes | 1.0 | 1.0 |
| | TFIDF | 0.6 | 0.75 |
| | KNN | 0.75 | 0.75 |
| | PRIND | 1.0 | 0.5 |
| Sports | Naïve Bayes | 0.71 | 1 |
| | TFIDF | 0.63 | 0.83 |
| | KNN | 1.0 | 1.0 |
| | PRIND | 0.6 | 1.0 |
| Baseball | Naïve Bayes | 1 | 0.67 |
| | TFIDF | 1.0 | 0.33 |
| | KNN | 1.0 | 0.67 |
| | PRIND | 0.5 | 0.5 |
| Soccer | Naïve Bayes | 1.0 | 0.75 |
| | TFIDF | 1.0 | 0.5 |
| | KNN | 0.8 | 1.0 |
| | PRIND | 0.75 | 0.75 |

Table 3: Ontology Classification Statistics

the dataset of 418 ontologies as a good representative of the entire ontology population.

We have performed a series of experiments to determine the most suitable algorithms for the ontology classification. For this purpose, we selected four popular classification algorithms - Naïve Bayes, TFIDF, KNN and PRIND. For the testing dataset, 22 ontologies were selected from five overlapping domains of interest: Sports, Baseball, Soccer, University, and Computer Science. The classification accuracy of the four classification algorithms was measured using the ontologies. Table 3 shows the relevant statistics obtained.

For evaluation purposes, we use True Positives (TP - Correctly classified), True Negatives (TN - Correctly unclassified), False Positives (FP - Incorrectly classified), False Negatives (FN - Missed classification). We compute the precision ($PC = TP/(TP + FP)$) and the Recall ($RC = TP/(TP + FN)$). Low recall means that our algorithm is missing many classifications that it should report according to our human expert. If the recall becomes 1, then there are no False Negatives. In that case, everything that should be reported is reported. Low precision means that our algorithm is reporting many classifications that it should not report. If the precision becomes 1, then there are no False Positives. In that case, only what should be reported is reported.

The experimental results show that Naïve Bayes is the most suitable algorithm for the OntoKhoj classification implementation. Also, it is interesting to see overlapping on-

tologies were correctly classified to a certain degree. Considering an example, *CS Department* ontology must be classified as *Computer Science* domain rather than *University* domain.

For the given classified ontologies, the OntoRank algorithm subsequently ranks them in a descending order. For performing experiments with the proposed ranking algorithm, we obtained 10 ontologies in Tourism domain through the OntoKhoj search interface. A subsequent execution of the algorithm on the dataset yielded results - a ranking of 10 tourism ontologies. A subjective evaluation of the results confirmed the correctness of the OntoRank algorithm. We admit that the subjective interpretation of our results is limited. Our research in the similar direction [7] focused on the development of metric-based ontology ranking method considering preferences of users. In future, we would like to incorporate a dynamic approach, wherein agents or users would express their own preference through the OntoKhoj portal for ranking ontologies. We foresee that a user oriented mechanism for ontology ranking would help in improving the practical accuracy of the results.

6. CONCLUSION

Responding to the compelling requirements of the Semantic Web community, we developed the OntoKhoj portal, which assists both humans and agents by simplifying the process of Ontology Engineering. The OntoKhoj development is based on novel methodologies allowing advanced searching, ranking, aggregating and classification of ontologies crawled from the Semantic Web. We focused on developing a proof-of-concept prototype of the proposed models and testing it on real Semantic Web data. Our claims are supported by experimental results of ontology crawling, ranking and classification, carried out with ontology data obtained from the Semantic Web. We believe that our OntoKhoj Web portal will provide knowledge engineers and agents a source for authoritative and trustworthy ontologies on Semantic Web, and expedite the process of Ontology Engineering through extensive reuse of ontologies.

7. REFERENCES

- [1] T. Berners-Lee, J. Hendler, O. Lassila, The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. In *Scientific American*, Mai 2001.
- [2] E. Garfield, *Citation Indexing: Its Theory and Application in Science, Technology, and Humanities*, John Wiley & Sons, New York, 1979.
- [3] S. Lawrence, C. L. Giles, K. Bollacker, *Digital Libraries and Autonomous Citation Indexing*, IEEE Computer, Volume 32, Number 6, pp. 67-71, 1999.
- [4] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, K. Miller; "Introduction to WordNet: An On-Line Lexical Database", <http://www.cosgi.princeton.edu/wn>, 1993.
- [5] Ontolingua Website, <http://www.ksl.stanford.edu/software/ontolingua/>
- [6] L. Page, S. Brin, R. Motwani, and T. Winograd, *The PageRank Citation Ranking: Bringing Order to the Web*, Stanford Digital Libraries Working Paper, 1998.
- [7] K. Supekar, C. Patel, Y. Lee, *Characterizing Quality of Knowledge on Semantic Web*, University of Missouri - Kansas City, Technical Report, 2003.