A Semantic Web Approach for Improving Ranking Model of Web Document

G.Charles Babu
Professor, Dept. of CSE
VVIT,Chevella

PVRD Prasada Rao
Professor, Dept. of CSE
PVPSIT, Vijayawada

N.Sandhya
Associate Professor, Dept. of CSE
GRIET,Hyderabad

V.Sujatha
Assistant Professor, Dept. of CSE
GITAM,Hyderabad

Dr A Govardhan
Professor & Principal
JNTU,Jagithyala

ABSTRACT

Ranking models are used by Web search engines to answer user queries based on key words. Traditionally ranking models are based on a static snapshot of the Web graph, which is basically the link structure of the Web documents. The visitor's browsing activities is directly related to importance of the document. However in this traditional static model the document importance on account of interactive browsing is neglected. Thus this model lacks the ability of taking advantage of user interaction for document ranking. In this paper we propose a model based on semantic web to improve the local ranking of the Web documents. This model works on Ant Colony algorithm to enable the Web servers to interact with Web surfers and thus improve the local ranking of Web documents. The local ranking then can be used to generate the global Web ranking.

Keywords

Web mining, Ranking algorithm

1. INTRODUCTION

In today's ICT era, information seeking has become a part of social behavior. With the plethora of information available on web, an efficient mechanism for information retrieval is primal importance. The search engines are an important tool for finding information on web. All of the search engines try to retrieve data based on the ranking of web documents. However the traditional ranking models on static snapshot of the Web graph, including the document content and the information conveyed. An important missing part of this static model is that the information based on user interactive browsing is not accounted. This affects the relevancy and importance metrics of a document, as the judgments of users collected by user at run time can be very important. In this paper we propose a model based in semantic web that enables a web server to record Interactive experience of user. The given approach works on three levels:

1. To make the existing model flexible i.e. the metrics related to relevance and importance can be modified according to run/browsing time user judgments.

2. The new enhancement should be automated in processing of the metrics recorded during browsing time.

3. The model should enable web server to play active role in the user's choice for highly ranked pages.

A Semantic Web Approach for Improving Ranking Model of Web Documents Our model has two components:

1. An ontology that keeps the interactive experience of user in machine understandable form.
2. A processing module based on Ant algorithm. Preliminary experiments have shown encouraging results in the improvement of local document ranking. The following sections give details about Semantic Web and various related terminologies afterwards there is brief discussion about Ant algorithm followed by proposed approach and experiment results with conclusion

2. Semantic Web

Semantic Web is an evolving extension of the Web in which the semantics of information and service on the web is defined [Lee 2007], which enables information not only to be process able by people but as well as machine. At its core the semantic web framework compromises a set of design standards and technologies. A typical RDF triple in ontology from proposed approach The formal specifications for information representation in Semantic Web are Resource Description Framework, a metadata model for modeling information through a variety of syntax formats and Web Ontology Language: OWL. The RDF metadata model is based upon the idea of making statements about Web resources in the form of subject-predicate-object expressions called RDF triple. Figure 1 provide a formal description of concepts, terms and relationships within a given knowledge domain.

The Semantic Web is based on a vision of Tim Berners-Lee, the inventor of the WWW. The great success of the current WWW leads to a new challenge: a huge amount of data is interpretable by humans only; machine support is limited. Berners- Lee suggests enriching the Web by machine-process able information which supports the user in his tasks. For instance, today's search engines are already quite powerful, but still return too often too large or inadequate lists of hits. Machine- process able information can point the search engine to the relevant pages and can thus improve both precision and recall. For instance, it is today almost impossible to retrieve information with a keyword search when the information is spread over several pages. The process of building the Semantic Web is today still heavily going on. Its structure has to be defined, and this structure has then to be filled with life. In order to make this task feasible, one should start with the simpler tasks first.

The following steps show the direction where the Semantic Web is heading:

1. Providing a common syntax for machine understandable statements.

2. Establishing common vocabularies.

3. Agreeing on a logical language.

4. Using the language for exchanging proofs.

Berners – Lee suggested a layer structure for the Semantic Web :

(i) Unicode / URI
(ii) XML/Name Spaces / XML Schema
(iii) RDF/RDF Schema
(iv) Ontology Vocabulary
(v) Logic
(vi) Proof
(vii) Trust

This structure reflects the steps listed above. It follows the understanding that each step alone will already provide added value, so that the Semantic Web can be realized in an incremental fashion. On the first two layers, a common syntax is provided. *Uniform resource identifiers (URIs)* provide a standard way to refer to entities,2 while *Unicode* is a standard for exchanging symbols. The *Extensible Markup Language (XML)* fixes a notation for describing labeled trees, and XML Schema allows defining grammars for valid XML documents. XML documents can refer to different *namespaces* to make explicit the context (and therefore meaning) of different tags. The formalizations on these two layers are nowadays widely accepted, and the number of XML documents is increasing rapidly. The *Resource Description Framework (RDF)* can be seen as the first layer which is part of the Semantic Web. According to the W3C recommendation RDF "is a foundation for processing metadata; it provides interoperability between applications that exchange machine Discovery of Semantic Web. Using Web Mining understandable information on the Web." RDF documents consist of three types of entities: resources, properties, and statements.

Today the Semantic Web community considers these levels rather as one single level as most ontology allow for logical axioms. Following ontology is "an explicit formalization of a shared understanding of a conceptualization". This high-level definition is realized differently by different research communities. However, most of them have a certain understanding in common, as most of them include a set of concepts, a hierarchy on them, and relations between concepts. Most of them also include axioms in some specific logic. To give a flavor, we present here just the core of our own definition, as it is reflected by the Karlsruhe Ontology framework KAON.3 It is built in a modular way, so that different needs can be fulfilled by combining parts. The Relation between the WWW, Relational Metadata, and Ontologies.

Definition 1 *A* core ontology with axioms *is a tuple* O :=(C;_C;R; _;_R;A) *consisting of _ two disjoint sets* C *and* R *whose elements are called* concept identifiers *and* relation identifiers, *resp. A partial order* _C *on* C, *called* concept hierarchy *or* taxonomy,*_ a function* _:R ! C+ *called* signature *(where* C+ *is the set of all finite tuples of elements in* C*),_ a partial order* _R *on* R, *called* relation hierarchy, *where* r1 _R r2 *implies* j_(r1)j = j_(r2)j *and* _i(_(r1)) _C _i(_(r2)), *for each* 1 _ i _ j_(r1)j, *with _i being the projection on the* i*th component, and _ a set* A *of logical axioms in some logical language* L.

This definition constitutes a core structure that is quite straightforward, well-agreed upon, and that may easily be mapped onto most existing ontology representation languages. Step by step the definition can be extended by taking into account axioms, lexicons, and knowledge bases [1].As an example, have a look at the top of Figure 1. The set C of concepts is the set fTop, Project, Person, Researcher, Literalg, and the concept hierarchy _C is indicated by the arrows with a bold head. The set R of relations is the set fworks-in, researcher, cooperates with, nameg. The relation 'works in' has (Person, Project) as signature, the relation 'name' has (Person, Literal) as signature.4 In this example, the hierarchy on the relations is flat, i. e., _R is just the identity relation. For an example of a non-flat relation, have a look at root.

Accommodation facility
Food Provider
Hotel
Minigolf
Youth -hostel
Belongs-to
Tennis – court
Family-hotel Wellness hotel
Is sports
-facility
Restaurant **Fast food**
Italian german
Fig 1 : Non – Flat Relation

Vegetarian - only regular Parts of the ontology of the content The objects of the metadata level can be seen as instances of the ontology concepts. For example, 'URI-SWMining' is an instance of the concept 'Project', and thus by inheritance also of the concept 'Top'. Up to here, RDF Schema would be sufficient for formalizing the ontology. Often ontologies contain also logical axioms. By applying logical deduction, one can then infer new knowledge from the information which is stated implicitly. The axiom in Figure 1 states for instance that the 'cooperates-with' relation is symmetric. From it, one can logically infer that the person addressed by 'URI-AHO' is cooperating with the person addressed by 'URI-GST' (and not only the other way around). A priori, any knowledge representation mechanism5 can play the role of a Semantic Web language. *Frame Logic* (or *F-Logic*; [5]), for instance, provides a semantically founded knowledge representation based on the frame and slot metaphor. Probably the most popular framework at the moment are Description Logics (DL). DLs are subsets of first order logic which aim at being as expressive as possible while still being decidable. The description logic SHIQ provides the basis for the web language DAML+OIL.6 Its latest version is currently established by the W3C Web Ontology Working Group (WebOnt)7 under the name OWL. Several tools are in use for the creation and maintenance of ontologies and metadata, as well as for reasoning within them. Our group has developed *Onto Edit* [3, 4], an ontology editor which is connected to *Onto broker* [1], an inference engine for F-Logic. It provides means for semantic based query handling over distributed resources. In this paper, we will focus our interest on the XML, RDF, ontology and logic layers.

3 TERMINOLOGIES

A Ontology is an explicit and formal specification of a conceptualization [Antoniou and Harmelen, 2008]. Ontology describes formally a domain of discourse. . It is a formal representation of a set of concepts within a domain and the relationships between those concepts. Typically, ontology consists of a finite list of terms and the relationships between these terms. The terms denote important concepts (classes of objects) of the domain. The relationships typically include hierarchies of classes. See Figure 1 where has Importance is a relationship between two concepts "document" and "importance". The "document" and "x" are the instances of these concepts. The major advantage of ontologies is that they support semantic interoperability and hence provide a shared understanding of concepts. Ontologies can be developed using data models like RDF, OWL.

…/system-document/world/document
…/system-document/world/document 1/importance
“document1”
“X”
Has Importance Resource ID
A Semantic Web Approach for Improving Ranking Model of Web Documents

B **Owl** The Web Ontology Language (OWL) is a family of knowledge representation languages for authoring ontologies, and is endorsed by the World Wide Web Consortium [Dean *et al*., 2004].

The data described by OWL ontology is interpreted as a set of "individuals" and a set of "property assertions" which relate these individuals to each other. OWL ontology consists of a set of axioms which place constraints on sets of individuals (called "classes") and the types of relationships permitted between them. These axioms provide semantics by allowing systems to infer additional information based on the data explicitly provided [Baader *et al*., 2003].

B Ant Colonies

Ant colonies are a highly distributed and structured social organization [Dorigo *et al*., 1991]. On account of this structure these colonies can perform complex tasks, which has formed the basis of various models for the design of algorithms for optimization and distributed control problems. Several aspects of ant colonies have inspired different ant algorithms suited for different purposes. These kinds of algorithms are good for dealing with distributed problems. One of these is Ant Colony, which works on the principle of ants' coordination by depositing a chemical on the ground. These chemicals are called pheromones and they are used for marking paths in the ground, which increases the probability that other ants will follow the same path

The functioning of an ACO algorithm can be summarized as follows. A set of computational concurrent and asynchronous agents (a colony of ants) moves through paths looking for food. Whenever an ant encounters an obstacle it moves either left or right based on a decision policy. The decision policy is based on two parameters, called trails and attractiveness. The trail refers to the pheromones deposited by preceding ants and the ant following that path increases the attractiveness of that path by depositing more pheromones in that path. Each path's attractiveness decreases with time as the trail evaporates (update) [Colorni et al., 1991]. With more and more ant following the shortest path to food the pheromone trail of that path keeps increasing and hence shortest optimal path is fund.

C Proposed Model

In our model we emulate the web surfing with the Ant Colony model. The *pheromone* counter of the links represents the attractiveness of the path to the desired Web document in our model. The more number of hits, the more important the link is but at the same time the hits are necessary to maintain the levels of importance the pheromone counter dwindles with time.

The Web surfers are the ants that navigate through the links of the Web documents to go to particular information. The Web server in this model is not a passive listener to cater the request of users but it is also the maintaining agent who records the pheromone of web links and ensures that updation of the pheromone is taken care of by the processing module.

4 MODEL WORKING

People looking for information visits web page through various links/page. Every visit is converted by the server into pheromone count and recorded. So if the person doesn't find the page useful he/she will not visit that page again and the pheromone count of that page will A Semantic Web Approach for Improving Ranking Model of Web Documents ¨ dwindle with time reducing it's attractiveness. Whereas repeated visits will increase the attractiveness of that page by increasing pheromone count. The web server records the pheromone count and also other interactive counts (more detail in the following section of server side enhancement).

A Server Side Enhancement

The server maintains the pheromone count and other interaction corresponding to a page in ontology and also periodically updates them. Sample ontology in Figure 2 describes the sample ontology. Currently we have two interaction metric recorded in the ontology: 1. Number of hits. 2. Visitor evaluation (1 = informative or 0 = not informative) relevance of the page. 3. Time Stamp (last visit)

```
< owl:Class rdf:ID=”hits”>
<rdfs:rdf:resource = “h-200”/>
</owl:Class >
<owl:Class rdf:ID = “ evaluation “>
<rdfs:rdf:resource = “e-1”/>
<owl:Class rdf : ID = “Date”>
<rdfs:subClasOf>
<owl:Class rdf:ID = “time-stamp”/>
</rdfs:subClassof>
</owl:Class>
<owl:Class rdf:ID = “time-hr”>
<rdfs:subClassOf rdf:resource=”#time-stamp”/>
</owl:Class>
<Date rdf:ID=”d-11-7-2008”/>
<time-hr rdf:ID=”t-1320”/>
```

Fig 2 : Server ontology

The above ontology shows that the document was visited on 11. July.2010 at 1:20 pm and it was the 200th hit.

B Pheromone Representation

Now we can sum up the entire picture by representing how the pheromone count works. The role of the pheromone is to record the trail and thus indicates the importance of the link/document. The count is always changing based on the time stamp last visited (i.e. the time elapsed after last count change by the visit).

So here is the how it works: The pheromone associated with the link/model is defined as: Pcount: D {V, T} (1) 50 ¨ A Semantic Web Approach for Improving Ranking Model of Web Documents Where V is pheromone density at a particular time and T is time stamp of last visit. Now the value of v can be updated in two ways: Positive update: When the user visits the page, user input of evaluation of page (positive) Negative update: With time the negative update decreases the pheromone count (evaporation). Also user input of evaluation of page (negative)It may be noted that equal weight age is given to user visit and input to avoid malicious degradation of pheromone so that the visit will cancel the malicious input. Say for example a user repeatedly visits a page and give negative input 0 but the updation account for the visit also so net negative updation is 0 but the evaporation will continue to take place with the last time stamp count of pheromone. The pheromone accumulation of a page at n+1 visit is done as follows:

Pnew = Pcurrent + 1 (2) The negative pheromone mechanism is realized by using the radioactive degradation formula: Pcount (t) = Pcount (T) * (1/2) exp (t - Pcount (T)/ (3) is the degradation parameter set heuristically. T is the last time stamp of updation so Pcount at time "t" is dependent on Pcount at last updation.

5 EXPERIMENTAL RESULTS AND CONCLUSION

We used the following model to ascertain the local page rank on a server setup using Apache Tomcat 5.5 for a collection of 70 web documents. Table 1 and 2 shows the observed result.

Table 1: Result for = 2

Percentage of page ranked within error margin of 10% 48*
Percentage of page ranked within error margin of 25% 56*
Percentage of page ranked within error margin of 40% 73*

Table 2: Result for = 4

Percentage of page ranked within error margin of 10% 21*
Percentage of page ranked within error margin of 25% 62*
Percentage of page ranked within error margin of 40% 87*

* Result value is approximated

The results clearly show the potential of this model. More than 50% of the pages were ranked within the error margin of 25%, which is encouraging in view of the sandbox environment of the experiment. In this paper we have presented our idea based on Ant Colony algorithm in the context of learning and web data mining. The proposed model proof of concept implementation shows the improvements in the existing system. The future work also holds promises in the area of improvement of current algorithm of Ant Colony by making it more relevant based on the information gain of the user experience.

Also other optimization can be in the fine-tuning of parameter and increment strategy of pheromone accumulation. We expect better results with more fine-tuning of the approach in future. A Semantic Web Approach for Improving Ranking Model of Web Documents ¨

6 CONCLUSIONS AND OUTLOOK

In this paper, we have studied the combination of the two fastdeveloping research areas Semantic Web and Web Mining, especially usage mining.

We discussed how Semantic Web Usage Mining can improve the results of 'classical' usage mining by exploiting the new semantic structures in the Web; and how the construction of the Semantic Web can make use of Web Mining techniques. A truly semantic understanding of Web usage needs to take into account not only the information stored in server logs, but also the *meaning* that is constituted by the sets and sequences of Web page accesses. The examples provided show the potential benefits of further research in this integration attempt.

One important focus is to make search engines and other programs able to better understand the content of Web pages and sites. This is reflected in the wealth of research efforts that model pages in terms of an ontology of the content. Overall, three important directions for further interdisciplinary cooperation between mining and application experts in Semantic Web Usage Mining have been identified:

1. the development of ontology's of complex behavior

2. the deployment of these ontology's in Semantic Web description and mining tools and

3. continued research into methods and tools that allow the integration of both experts' and users' background knowledge into the mining cycle. Web mining methods should increasingly treat content, structure, and usage in
an integrated fashion in iterated cycles of *extracting* and *utilizing* semantics, to be able to understand and (re)shape the Web.

REFERENCES

[1] Antoniou and Harmelen, A Semantic Web Primer, pp. 11, MIT Press, Cambridge, Massachusetts, 2008.
[2] Dorigo, Maniezzo and Colorni, The ant system: an autocatalytic optimizing process, Technical Report TR91-016, Politecnico di Milano, 1991.
[3] Dean, W3C reference on OWL, W3C document, 2004
[4] Dorigo ,A.Colorni and V. Maniezzo, Distributed optimization by ant colonies, Proceedings of ECAL'91, European Conference on Artificial Life, Elsevier Publishing, Amsterdam, 1991.
[5]Lee, MIT Technology Review, 2007