Link Annotation

Bilal Gonen
Maciej Janik
Samir Tartir
Ravi Pavagada

Department of Computer Science, University of Georgia
Athens, GA 30602
{gonen, mjanik, startir, ravipr}@uga.edu

1. Introduction

The network of hyperlinked documents, as it exists now, lacks semantic information in machine understandable form. It can only be browsed or searched by keywords - not concepts. There do already exist projects that automatically or semi-automatically annotate web pages with concepts taken from ontology. This effort makes web pages more understandable for machine processing and searching. In our project we would like to focus more on navigational implications of adding semantic annotation to web pages.

Currently user or machine navigates between web pages by traversing them via hyper links. Decision if accessed page is relevant to the undertaken search can be made only after retrieving and analyzing the destination web page. In our project we would like to add more semantic meaning to links themselves on the source page, so concepts included on target page can be evaluated without retrieving page itself.

Network of web pages

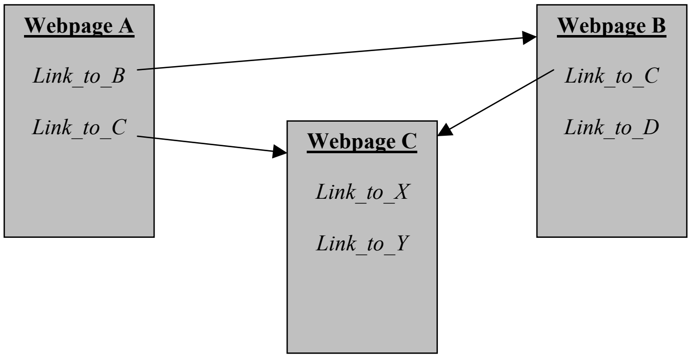

Figure 1 - Network of web pages

The general idea of link annotation is already known and called MREF (metadata reference link). We would like to implement MREF in practice in the field of semantic web. This enhancement will enable better navigation between pages, as target concepts are known at browse time of one page.

]

System Architecture - Overview

We would like to make our system modular and expandable for future needs. As we cannot modify the content of web pages, we can only keep discovered annotations of pages and links in snapshot of selected web pages.

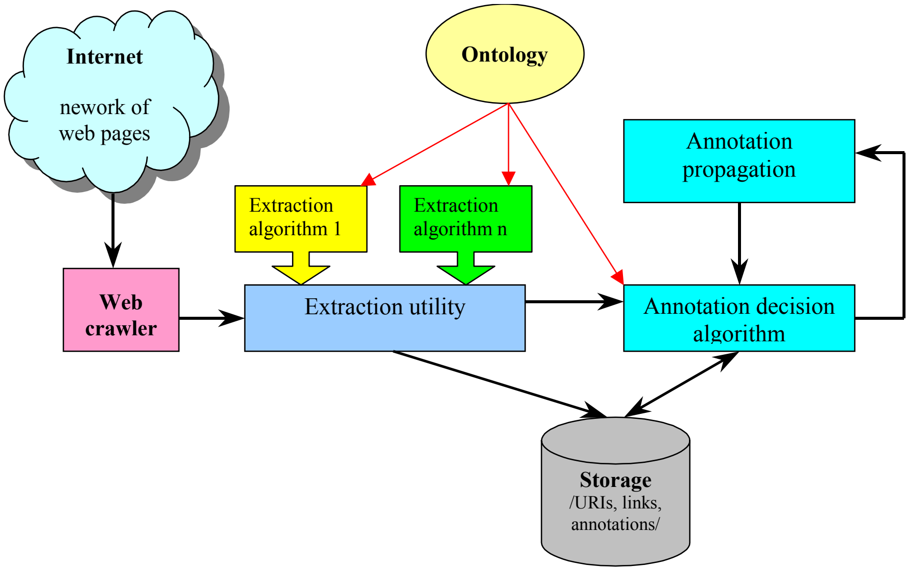

Figure 2 General system architecture

First module is a WebCrawler that crawls the web structure and supplies the raw data for further analysis. Than HTML from web pages is analyzed by extraction utility. Here extraction mechanism tries to match the whole page to some concepts in ontology. It also tries to categorize links in this web page based only on information contained in this web page (not taking the target page into consideration). All extracted information is stored in persistent storage.

As last, came the decision and propagation loop. Now the extracted information is analyzed again. Some of extracted information may be deleted from page, some can be inferred or pushed from links to page. In this step web pages are analysed in network and we allow annotation flow between nodes. Both from page to describing link and from link to described page. This is iterative process and after few iteration the network should reach some stable (or near-to-stable) state.

In such state we would say that the selected network is annotated and can be used in semantic navigation.

Related Work & Our Ideas:

There are lots of papers on HTML Tag tree extraction or deriving link context, one of them is “Deriving link-context from HTML tag tree” by Gautam Pant et al. There are lots of other papers on automated semantic annotations; among them is a paper on “SemTag & Seeker: Bootstrapping the semantic web via automated semantic annotations”. We would like use some of the ideas from “Mining the link structure of semantic web”, by Souman Chakrabarti et al., which talks about HITS algorithm. The algorithm takes advantage of the hubs in some fields & uses techniques that take advantage of social organizations of the web. It allocates weights for the hub pages &

authorities in iterative process. The paper “On extracting link information by relationships instances from a website” by Myo-Myo Naing et al, talks about a web page which is being associated with a concept in ontology & links two different web pages based on the relationship between concepts in the ontology. We would like to incorporate this in our project. The authors of this paper do not consider intermediate pages whose relationship is not defined in the ontology. This paper is solely based on concepts & their relationships described in ontology. We would not only like to define relationships between web pages based on ontology, but also look at the link context & extract the relationship information. Concept matching is an area in itself & there are lots of papers on it. There are lots of AI & natural language processing techniques used to achieve this. We would like concentrate more on the information which is around the link, i.e. link context & match the link to a concept. We would also like to look at the contents of the linked page & match it to a concept. In most cases there might be more than one concept matching for a given webpage. This is something we need to put in more thought on. Annotated information is propagated from the parent webpage to its linked webpage & also from the linked webpage to the parent webpage. We will be annotating both the link webpage & the parent webpage based on some relationship & the matching concept.

Approach:

The approach will work in two general phases: Preparation and Annotation.

I. Preparation:

Here a deep analysis of the Computer Science department in the University of Georgia will be conducted, resulting in building an ontology that represents the current structure of the department.

This resulting ontology will be used in the next phase for annotation.

II. Annotation:

This is where the actual process of page and link annotation will take place. This phase is divided into three stages:

1. Page annotation.
2. Link annotation.
3. Relationship annotation.

1. <u>Page annotation:</u>

In this stage, all the pages in the Computer Science department site will be analyzed in one of the current methods, or a new method that we might need to develop. The result of this analysis will be a mapping between a certain page, and a node in the ontology designed in phase I.

2. <u>Link annotation:</u>

Here, each page will be scanned for links that point to pages in the same domain, and each link will carry the annotation of the page it points to.

3. <u>Relationship annotation:</u>

This is the final stage that defines which type of relationship the link defines. This relationship is obtained from the ontology based on the types (concepts) of the page with the link, and page the link points to.

The resulted annotated pages will be stored in a database the application has access to write to and issue queries against.

References

1. Effects of Link Annotations on Search Performance in Layered and Unlayered Hierarchically Organized Information Spaces
http://lhncbc.nlm.nih.gov/lhc/docs/published/2001/pub2001040.pdf

2. Mining the Link Structure of the WWW
http://citeseer.ist.psu.edu/chakrabarti99mining.html

3. Deriving Link-context from HTML tag tree
http://dollar.biz.uiowa.edu/~pant/Papers/tagTree_dmkd.pdf

4. Search Engine-Crawler Symbiosis: Adapting to Community interest
http://dollar.biz.uiowa.edu/~pant/Papers/se-crawler.pdf

5. Automatic resource compilation by analyzing hyperlink structure and associated text.
http://marco.uminho.pt/disciplinas/UCAN/BD/Artigos%20Recomendados/chakrabarti98automatic.pdf

6. On Extracting Link Information of Relationship Instances from a Web Site
http://citeseer.ist.psu.edu/662560.html

7. Semantic Blogging and Bibliography Management
http://www.w3.org/2001/sw/Europe/reports/open_demonstrators/hp-requirements-specification.html

8. Ontology-based Web Annotation Framework for Hyperlink Structures

9. A Study of User Model Based Link Annotation in Educational Hypermedia
http://www2.sis.pitt.edu/~peterb/papers/JUCS98.pdf

10. Semantic Linking - A context based approach to Interactivity in Hypermedia

11. Web Document Searching Using Enhanced Hyperlink Semantics Based on XML
http://www.db-net.aueb.gr/hercules/papers/ideas.pdf

12. Automatic Annotation of Content-Rich HTML Documents: Structural and Semantic Analysis
http://citeseer.ist.psu.edu/668883.html

13. Semtag and seeker: bootstrapping the semantic web via automated semantic annotation.
http://www.almaden.ibm.com/webfountain/resources/semtag.pdf

14. Media-independent correlation of Information: What? How? -- MREF paper
http://www.computer.org/conferences/meta96/sheth/