Web Document Searching Using Enhanced Hyperlink Semantics Based on XML[1]

I. Varlamis, M. Vazirgiannis
Dept of Informatics,
Athens University of Economics & Business,
Patision 76, 10434,
Athens, HELLAS
{varlamis,mvazirg}@aueb.gr

Abstract

We present a system that aims at increasing the flexibility and accuracy of information retrieval tasks in the World Wide Web. The system offers extended searching capabilities by enriching information related to hyperlinks between documents. It offers to documents' authors the ability to attach additional information to hyperlinks and also provides suggestions on the information to be attached. In an effort to increase the integrity of hyperlink information, a conversion module extracts, from the pages, metadata concerning the linked documents as well as the link itself. The hyperlink metadata is appended to the original document metadata and an XML document is created. Another module allows the end-users to query the XML-document base, taking advantage of the enhanced hyperlink information. We present the overview of the system and the solutions it provides in problems found to similar approaches.

1. Introduction

The amount of information stored in the World Wide Web continues to grow rapidly and the search engines' attempts to organize it into comprehensive indexes that satisfy a user's query are often inadequate, resulting on very high or low recall or in ambiguous precision of the retrieved results [1]. To enhance the retrieval mechanism of search engines and increase their flexibility in retrieval of relevant documents to user's query, additional information must be attached as well as new types of associations among the retrieved information pieces.

Another problem that search engines confront is that they mainly parse pages' contents automatically in a "blind" way, with limited human support. In this way the results to a user's query are a set of links to the pages the parsing algorithm considers to be of interest.

A feasible approach towards augmenting information space is to enrich information regarding the interconnectivity of web documents. Without doubt, people who have visited the page and moreover have created a reference to this page can provide more accurate knowledge on a page's contents. Therefore there is an emerging need for a link enhanced search engine. The results of such engine will be a set of URLs created and described by human authors in their own web pages that point directly to the destinations of interest. This set of "suggested" URLs would be of known source and destination concept as well as link type.

> ***Concept*** *is a set of keywords that characterize the documents' content.*
>
> ***Link type*** *is a set of words that describe links' semantics.*

As Lempel and Moran [2], who introduced the idea of informative links said, "*the hyperlink provides a positive critical assessment of destination page's contents, which originates from outside the control of the author of the destination page*". Lempel and Moran suggest the authority hyperlinks as recommendations from a web page author to other users. This contributes to the idea of creating an information model, based on how other users view or comment a web page's contents, through the hyperlinks they create towards it.

Previous approaches of exploiting link information can be divided into three categories:

a) Identification of link interconnection patterns over a set of pages,
b) Classification of links into categories based on the type of documents they connect,
c) Attachment of additional semantics to links describing the source, the destination, the type of link and a potential weight.

The latter approach derives from the open hypermedia research area and proves the need for extended information concerning the link properties. [3].

Obviously in the case of World Wide Web documents, hyperlinks have poor information content and lack of semantics although they represent logical relationships between the information nodes (web documents). Such

[1] This work has been supported by the GSRT/PENED #99EΔ85 project

relationships present special characteristics, like inheritance or directionality, which can be proved valuable if they are properly exploited. An appropriate semantic model can broaden the scope of web information searching, allowing flexible information selection that will assimilate the user's interests.

Given that we classify each web document into a conceptual category, we can simulate links as relationships between two concepts. For example we consider the following case:

> *A web page contains an abstract and a link to the full paper. Another web page contains a summary of a book and a link to the online version of the book.*

The first link can be defined as "abstract" to the linked model. The logical relation "The abstract ***abstracts*** the paper" contains a directionality that cannot be inverted, because a paper could never abstract an abstract. Similarly "The summary ***summarizes*** the book" and the relationship is uni-directional. In a rich semantic model that supports inheritance is easy to define that an abstract is ***kind of*** summary and a paper is ***kind of*** book. The inheritance between the concepts presumes an inheritance between relationships. As a consequence the relation "an abstract ***summarizes*** a paper" can be supported.

Such definitions are based on human decision. If we consider web as a large information source that covers many domains of interest the knowledge of domain experts is required for the definition of the relevant concepts and of their hierarchy. Hyperlink authors need a flexible and generic model that encapsulates the conceptual hierarchy knowledge and allows rich semantics to be attached to hyperlinks. This hierarchy of documents' concepts and the relations among them comprise a knowledge network that can be utilized to facilitate enhanced searching tasks.

Section 2 presents relevant work done in hyperlink information area. In section 3 the general architecture of the system is introduced and the technologies employed are discussed. Subsections discuss the various innovations and benefits the system induces, such as the taxonomy of links, the assignment of a degree of relevance to certain concepts and the use of XML-Schema for validating XML documents that contain enhanced hyperlink information. Section 4 describes the modules that comprise the system and a prototype that is being developed. Finally, section 5 provides some conclusions on the proposed model, and weighs against the solutions provided and the problems faced.

2. Related Work

A first attempt in exploiting link information drawn from a set of pages would be to maintain a database with the source and destination of each link. Many approaches that are based on link structure mainly focus on identifying patterns of link interconnection. Kleinberg [4] uses the number of incoming and outgoing links of a page to characterize the page as hub (many outgoing links) or authority (other pages refer to it). He employs iterative algorithms to find distinct communities of pages in his link database and considers that pages with similarity in outgoing links share the same subject. Spertus [5] uses link's source and destination information to create "Link Geometries", which are sets of pages with similar subjects. She also uses links interconnection to correct problems of moved pages, missing links etc. Both these, and other similar approaches [6], take advantage only from information stored in the <HREF> tag and ignore link content and concept information.

A better approach suggests extracting more information, concerning the link, from the document. Definitely it is difficult to perform a full lexical analysis in an HTML document and produce a meaningful summary. Although it is hard to extract all the concepts that an HTML document introduces, work that is done in parsers' and wrappers' area as well as recent work in document clustering [7], [8] proves that it is fairly simple and accurate to extract the general concept of a document. The former attempts extract semantic features from the words or structure of an HTML document. These features can be assigned to the whole document or to a specific part of it, for example to a link. The latter ones discover clusters of documents that share the same concept or concepts and set the criteria for assigning new pages into a cluster. Thereby a link or a page obtains content that can be used to classify and categorize it in broad hierarchy, where page concepts are mapped to nodes and links between pages to relations between concepts. The results of such categorization could be used to formulate queries that profit from conceptual relations. The knowledge network created is an abstract view of the web itself. This abstraction seems to be necessary in order to handle the huge amount of information contained in the web.

There are many works that try to define different categories of links and classify new hyperlinks in these predefined categories. Such works were mainly focused on extracting conceptual links from unstructured documents (text documents) and not on characterizing already existing links. Allan [9] presented methods for gathering documents for a hypertext, linking the set and annotating the type of the link. He distinguishes three types of links according to whether they can be retrieved automatically or based on pattern-matching or other techniques. However, the automatic techniques introduced can be applied to the link description effort.

The techniques used by Allan employ the comparison of terms in a paragraph or sentence level and the creation of a detailed link graph for one or more documents. By simplifying the graph (either by grouping a number of similar links to a single link, or by increasing the similarity threshold) one can set topic boundaries inside a document and generate conceptual links between them.

Allan's effort was based on a previous taxonomy performed by Trigg [10], who lists a set of 80 classes of link types, some of which are: revision, summary and expansion, equivalence, comparison and contrast, tangent, aggregate. Actually this classification concerns the paper readers and authors community, thus the classes are defined to cover the concepts of interest to this community.

HTML 4.0 approach attempts to describe link types by defining a limited number of link types for the web pages' creators. Some examples are "Next" and "Prev" links, "Chapter", "Section" and "Subsection"[11]

Another interesting work is the VOIR paradigm [12]. VOIR takes into account users' attitude in anchor selection in order to rate an anchor's relevance to a certain subject or to other anchors. Although Golovchinsky's work is not relevant to link characterization, it introduces a very important notion, the one of relative relevance of a link to a certain type.

This ambiguity of link participation in a certain concept, can be proved very helpful in information retrieval task. Petrou, Martakos and Hadjiefthymiades [3] presented a similar idea in hypermedia semantics area. They present the idea of a hypermedia layer, where information on link type and weight will be stored.

Although the idea of characterizing links, or adding semantics to links was discussed a lot, the weighting of links' relevance to a concept and its relevance to link types hierarchy, was not adequately covered. The need for such information is also stated in adaptive hypermedia models, where one of the link-adaptation techniques is the link annotation according to its relevance to a subject.

3. The proposed architecture

Currently the majority of links in web pages are simple references from a source to a destination node. Links mainly describe the target node and provide information concerning his location and his main concept but lack of information concerning the semantic characteristics of the association that the link designates. It is straightforward that by attaching semantic characteristics to the link it is possible to enhance the hypertext network to a knowledge network. It is as well important for web users to have concise information on the basic concepts of the linked page. The creator of the page will be responsible to provide such information.

A basic requirement is the categorization of the semantic relationships between web pages. With the use of XML it will be easy for the web page creator to precisely define the type of each link, and additionally to provide information about the entire taxonomy of concepts of interest. This information will help the search engines task. It is also possible to give a short description of the linked page, either as a thumbnail image [13] or as a group of descriptive terms. Phelps and Wilensky [14] support that five words are enough to uniquely identify a web page and describes a process of automatically extracting these words out of a web page, mainly based on term frequency and inverse document frequency.

In the current approach we propose a structure for the enhanced hyperlink semantics. We also propose a process of concept definition and categorization in a hierarchical structure. In the following we will discuss the general architecture of the system as this is depicted in figure 1.

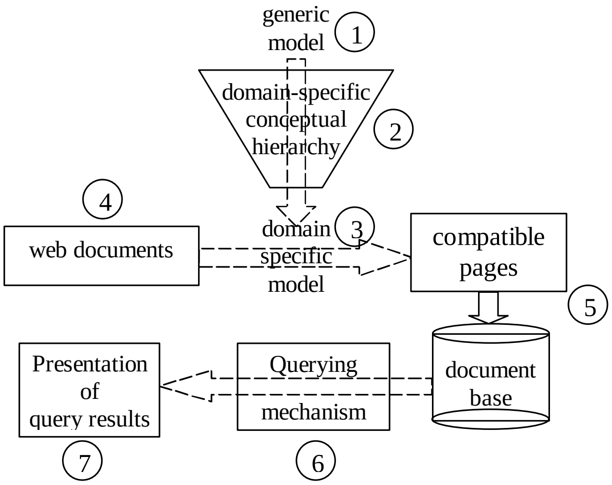


Fig.1 The general system architecture

The first step in our approach of exploiting rich hyperlink information is the definition of a generic structure for the information stored into the hyperlink (Fig 1, step 1). Web documents may cover a great range of concepts and may be addressed to different groups of interest. In each domain of interest a group of experts must define a conceptual hierarchy for documents and link types (Fig.1 step 2), which will be mapped into the hyperlink information model for the specific domain (step 3).

The next phase is the collection of information from web documents and the creation of new documents that contain the additional hyperlink knowledge structured according the domain specific model directives (Fig 1, steps 4 and 5). Considering these directives the system allows users to query the collection of documents (step 6) and returns a set of results for each query (step 7).

The additional link information can be used in many ways:

- Search engines will provide more advanced search features, since users can search for pages linked in a specific way.

- Results will be richer in knowledge, since information about the degree of relevance will be presented
- Search results will be ranked more precisely based on link information that comes from their neighboring pages.
- The search results can be easily grouped according to the conceptual classification or to the type of incoming or outgoing links they have.
- Conceptual hierarchy will provide new ways of transition between concepts in the resulting set of pages.
- Navigation among the various web pages will be more informative since browsers will be able to present the additional hyperlink information to the users. [13]

As stated in [15] the reader's benefit from link information is a well-described linked set of hypertexts that allows the creation of graphical maps and the ability to display link information to the user without the need to follow the link. Additional hyperlink information can be very useful in certain cases as guided tours and computer based training hypertexts, where navigation among pages follows a predefined path or changes according to user's preferences. With the support of a powerful query language the amount of available link information can be the next solution to the problem of navigation in the web. Current presentation techniques such as static site maps, clustering techniques and fish eye view can be replaced by the practical and transformable graphs that result from a query on the meta-information hidden inside the links.

3.1. Generic structure of hyperlink information

The current model consists of a hierarchy of classes that contain information on various aspects of a hyperlink (as shown in Figure 2). The nodes of hierarchy represent the classes of the model. The edges represent containment. The notation used in the graph is borrowed from the XML-DTD notation, where:

A* , *B *denotes that A consists of Bs*
A* , **B *denotes that A consists of 0 or more Bs*
A* , *1B *denotes that A consists of exactly 1 B*
A* , *+B *denotes that A consists of 1 or more Bs*

Symbol <|> denotes the logical OR in containment relation. The root object represents the source document and contains information that relates to its concepts. It also contains a list of URL objects. Each URL object contains information on itself as well as on the document it points to.

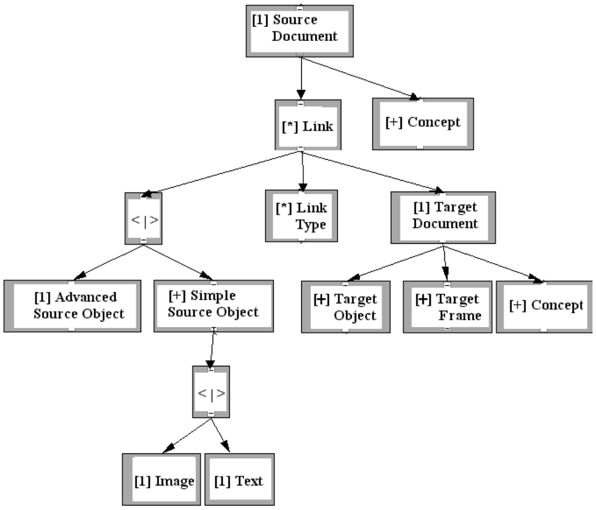


Fig.2 The hyperlink information structure

The hyperlink can be of one or more types and the target document is characterized by a list of concepts.

Concept class has two attributes: A string for the name of the concept or type and a percentage value representing the relevance of the characterization.

Advanced source object mainly refers to Java or ActiveX components that may contain link to other pages, HTML forms etc.

Link Type class stores the knowledge and uncertainty of the link. "Link Type" describes the kind of conceptual linking between the two nodes. An example could be that a link from a node A of type "paper title" to a node B of type "*abstract*" would establish a relationship of type "*is abstract*". The "relevance" object assigned to this link can have a percentage value denoting how much relevant is the link type. The link type description in this case would define certain restrictions on the type of nodes a connection can have. To illustrate this we can think of the rule: " a relation of type "*is_abstract*" can only be established between concepts of type *paper_title* and *abstract* and only towards the direction of *abstract*".

3.2. Domain-specific conceptual hierarchy

Since the information that resides in the web covers almost any concept, a first step will be to classify this knowledge into conceptual groups, further called concepts, probably based on ontology theory. On the conceptual level Davenport and Prusak's statement [16], "people can't share knowledge if they don't speak a common language", is utterly crucial for the case of web information.

Staab et al [17] discuss the role of ontologies in the area of intelligent information integration as information mediators between distributed and heterogeneous information sources and the applications that use these information sources.

Every group of users will be able to define a set of concepts that characterize its interests. This task is better to be performed from a group of experts in the current domain. The creator of a hyperlink from a web page to another page defines the group of concepts the linked page refers to and the degree of relevance to each concept. He also defines the concepts of the source page and the type of relation the link denotes.

As Berstein [18] states “a vocabulary of link types might prove useful in a variety of ways”:

- Readers could use link types to understand the structure of a document.
- Agents can assist the readers by interpreting the link types. They can change the way a link is presented according to his type.
- By providing a rich vocabulary of link types, the system could encourage writers to create richer and better-structured arguments.

Trigg [10] initially stated the problems that this classification can cause:

- The number of link types must be limited and static. If there are no restrictions, the system will not be able to handle every new type created by users.
- The reader and the system will be confused unless the creators of new types somehow define the type to the system.

This problem can be solved using XML. With XML, users are able but also obliged to define any new link type, as well as its position in the link types family.

XLink [19] is the linking language that allows elements to be inserted into XML documents to create and describe links between resources. XLink allows the separate storage of the link meta-information and the linked resources.

Although XLink provides a sturdy base for building explicitly linked documents, it does not fully cover the need for relevance to concepts and conceptual hierarchy. However, XLink’s flexible architecture can be expanded to cover such issues since it allows the creation of new elements that can describe both the destination subject(s) and the degree of belief in each subject.

Links between World Wide Web documents represent logical relations between web pages. By characterizing and weighting each link the expressive power of hyper-linked documents increases and a new query algebra emanates that permits retrieval of pages based on their relevance to a subject

The proposed system induces the notion of enriching the link with relevance semantics, in order to turn www network into a flexible tool for organization and manipulation of knowledge. Each link that connects two pages will be associated to one or more types with one or more degrees of belief respectively. To provide an example:

A link from a newspaper article concerning a company’s strategic movement, to the company’s ‘News’ page will be an “expansion” with a weight of 90. The link can also be classified as a “comparison” with a lower weight.

The above model can be extended to rate proximity relations. They do not contain specific semantics but denote a semantic proximity between the two linked documents. Two documents obtain a proximity relation when they are positioned adjacently in a logical frame.

By extending the above example, two financial articles that appear in a newspaper front-page and link to two different pages, establish a relationship between the target pages, which can be considered “close” to each other.

3.3. The XML-Schema

A main necessity for the model for being generic and able to cover the volume of link information in the web is the existence of a sturdy foundation. A template document must be defined that will allow easy and uniform management of hyperlink information that hides on the web documents.

The use of XML as the language for storing the hyperlink information induces many benefits:

- XML information has internal structure that follows a certain scheme, so it can be easily searched, queried or stored in a database schema.
- Emerging query languages are XML oriented and provide great capabilities such as: path navigation, result transformation, grouping, sorting and restructuring. [20]
- XML information can be stored within or outside the original html page. In the second case, many XML files can be produced from the same html source, since each search engine (or anyone interested in the page) can create an XML description of the page’s hyperlinks and keep it in its own XML database. Thus for an HTML page of general interest two communities can have two different XML documents with their own link semantics, from their own point of view.

A basic step on building our information model is to express it in a language that can be easily handled afterwards. The simplest approach was to describe the model by creating a DTD document [21]. DTDs provide an excellent set of tools for describing document structures, as well as a powerful set of tools for reusing descriptions. There is a strong relationship between DTDs and XML since they can be used to validate XML documents’ structure. Additionally, the model proposed can be easily and with a limited set of commands defined in a DTD document. Based on the model proposed in figure 2, we created a DTD document that can be used to express the proposed model as well as to validate the XML documents that will contain link information. In the DTD document, link types were set based on Trigg's work [10] that concerned bibliographic work. In order to cover different areas of interest, a different set of available link types must be defined, consequently a new DTD.

DTDs have been introduced by the document community as a means for specifying the document structure from the parser viewpoint rather than from the DBMS one [22]. As a consequence, there are several shortcomings:

- DTDs do not support inheritance between concepts, which is very crucial in our project.
- In the case of link information, different communities will have to use different DTDs to validate their XML link documents, using different link types or document concepts.
- The hyperlink information model described needs a percentage data type to express weight, and needs to set limitations in what kind of concepts a certain type of link can connect.

In order to bypass these limitations of DTDs we used XML Schemas to describe our model. A fragment of the XML Schema proposed, derived from the original DTD file, can be found in the appendix.

4. A prototype system

In order to test the abilities of the model a prototype system is developed aiming to create a collection of web hyperlink information and provide the means to query this data set.

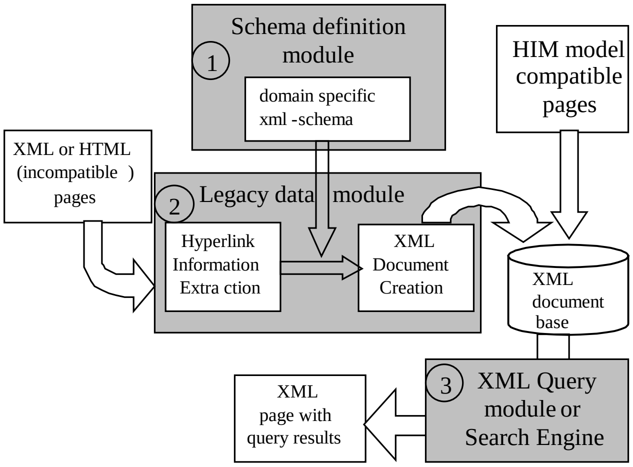

Schema definition module
1
domain specific xml -schema
HIM model compatible pages
XML or HTML (incompatible) pages
2
Legacy data module
Hyperlink Information Extra ction
XML Document Creation
XML document base
3
XML Query module or Search Engine
XML page with query results

Fig. 3 The architecture of the proposed system

The domain-specific conceptual hierarchy that was described in section 3.2 needs to be mapped into a set of rules that will control the validity of XML documents. This set of rules will be the XML-Schema for the documents of a certain domain of interest. Therefore domain or communities experts need a tool for creating domain-specific XML-Schema files. This is the schema definition module of our system (figure 3, part1).

The source of the hyperlink information is both HTML and XML pages. In the case of valid XML pages (pages that follow the XML-Schema rules) that contain hyperlink information, the URL of the document is stored in the document base. The legacy data module (figure 3, part 2) converts HTML pages or invalid XML pages to XML pages that follow the model directives. The validity is always tested against the appropriate XML-Schema for each domain. The valid XML files are stored locally and the document base is informed on their location. Search capabilities over the collection of XML documents are to be provided through a simple user interface where users will assemble their queries in an XML query language. Query languages as XML-QL offer the ability to reconstruct an XML document that contains the query results. The query module (figure 3, part 3) can be replaced or attached to a typical search engine to enhance or filter the search results exploiting the additional information.

4.1. Schema definition module

The hyperlink information model must be able to serve the needs of all possible communities of users. Thereby, it must provide the ability to groups to define their own concept and relationship descriptions. They must be able to define the concept types and organize them in a hierarchy starting from the more general concepts and branching down to the most specific ones. Relationships among the most general concepts can be inherited to their children, grace to the object oriented nature of the model. Therefore:

> *A relationship of type "advertises" between an "advertisement page" and the "advertised product" page can be inherited to the relation between "yellow pages" and the pages they link to.*

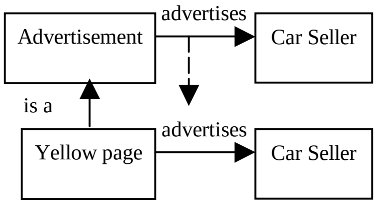


Fig. 4 A example of link inheritance

The definition of the available concepts and relationships and the construction of a hierarchy with allowable relationships between concepts is the first step in creating a proper information space. The definition can be included in a file or set of files (XML-Schema files) that will become available to users of each community.

Each web page developer can refer to the definition file (XML Schema) of his/her interest and provide pages with the necessary metadata that describe the concepts, relations and weights.

The concepts will be organized in a hierarchical order, thus permitting inheritance issues to operate. The taxonomy of link concepts is used to enhance the query process. By employing link types hierarchy, a search engine expands the search process in relevant areas of a subject in order to broaden the result set, or distinguishes a certain branch (according to user selection) and focus on a specific conceptual area.

Having defined this conceptual framework, it is easy for the web page developer to create a link and provide all the relevant metadata needed to describe the source and destination relevance to certain subjects.

If a search engine knows the conceptual hierarchy in advance, it should be possible to direct the search into pages with relevant subjects. For example:

> *When searching for information on people that know programming, we can search in the relevant areas of students or professors who know programming. In this case the search is based on the fact that concepts "student" and "professor" are specializations of the concept "people".*

In a similar way:

> *When searching for papers that relate to XML, the initial search can be done in the set of pages that contain papers but can also be expanded in the more general category of pages that contain books on XML. The latter results will be rated lower than the former.*

Grand Central Station [23] is a project that uses topic hierarchy to expand the relevant document base when performing a search in its database.

4.2 Legacy data module

The most tedious task for the system is the characterization of legacy web pages and hyperlinks.

A first simple approach will be to use terms from:

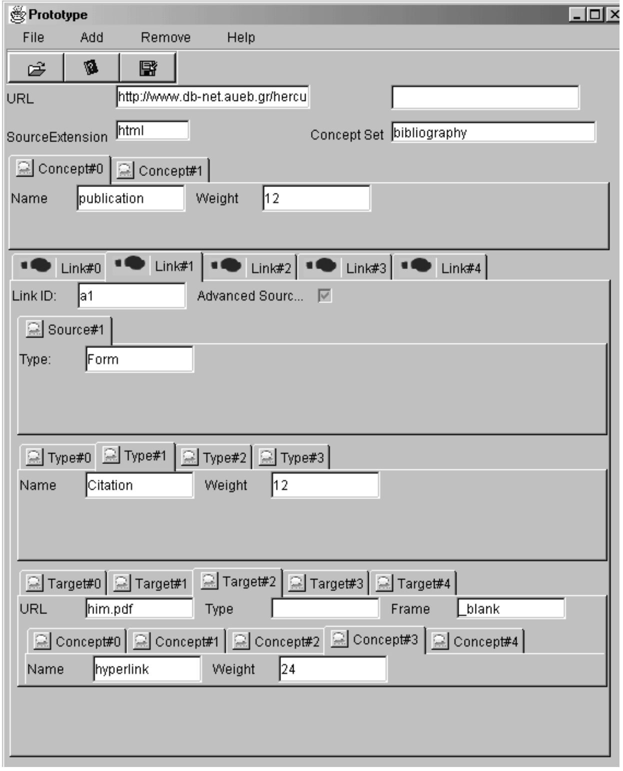
- Source document's title
- Destination document's title
- Link description
- The last heading before the link
- The hyperlink position

These techniques are already used in similar projects to characterize a page's contents [23], [24].

Mizuuchi et al [6] use the pages that lead to a certain page to enhance with keywords the pages' contents. From the contents of the "entrance path" pages they search for keywords in tags <title>, <a> and <h1> to <h5>.

In order to support this process a tool is built that handles the information concerning the links an html page contains. The basis for this tool is an html parser written in Java, which extracts information that relate to the hyperlinks of an html page and presents them to the user.

The information that is extracted is adapted to the requirements of the XML-schema described in section 3.3. Through a simple user interface the user can filter and modify the information presented. A snapshot of the user interface is presented in figure 5.



URL http://www.db-net.aueb.gr/hercu

SourceExtension html

Concept Set bibliography

Name publication Weight 12

Link ID: a1 Advanced Sourc... ☑

Type: Form

Name Citation Weight 12

URL him.pdf Type Frame _blank

Name hyperlink Weight 24

Fig. 5 The user interface of the legacy data transformation module

Since the user selects an html document (either local or remote) the document's contents are parsed and a set of tabbed cards is created. Each card refers to a link and contains a set of fields that contain extracted information that concerns the link. The user may characterize the link type, the source and destination page type, and assign weights. It is also possible to modify the information already extracted by adding new concepts or removing other that are not relevant or comprehensive. The available concepts or link types each time depend on the selected conceptual set.

Once the user has finished editing the properties for each link of a certain html page, he/she can save the additional link information to a new XML document that is compliant to the XML-schema. The XML document may either contain only the link information or embed it in the original page content, thus creating an XML page that can substitute the original HTML page. The first solution is appropriate when the user does not have the privilege to change the contents of the web page. The second solution will be very helpful for web administrators who want to convert their HTML-pages into XML-pages with enhanced link information.

4.3 The query module

In order to make use of the additional hyperlink information that the model suggests, three issues are important:

- Selection of an appropriate query language that will be able to cover user needs and additionally provide advance capabilities on the result set such as sorting, aggregation, restructuring etc.
- Design of a smart, practical interface that will help even naïve users search for subjects of interest by fully exploiting the information network capabilities.
- Definition of a visualization model that will present most of the available information in a conceptive manner and will allow user intervention on the query results.

Regarding the query language, an XML-like query language that supports the basic functionalities, as those described in [25], is the best solution.

To demonstrate the advantages of the use of rich hyperlink information we use XML-QL query language [26] to express a few of the possible queries that can be addressed to the information base. Of course other XML query languages as Quilt [27] or techniques presented in [28] can be applied to the problem.

The following query will produce an XML document that contains links to Doe’s page as those are stated in Kobe’s University pages

```
    where <Document*>
    <SourceDocument>
    <Concept Name=”Homepage” and “Kobe
  University”></></>
    <Link><TargetDocument>
    <Concept Name=”Homepage” and “Doe”></>
    <TargetObject URL=$U> </></>
    <LinkType Name=”User Homepage”></> </></>
  in www.a.b.c/hyperlinks.xml
    construct <link>$U</link>
```

The above XML-QL query is applied to a single valid XML document (namely *www.a.b.c/hyperlinks.xml*). Starting from the root element (<Document>), the query seeks for SourceDocuments of certain Concept (“Homepage” and “Kobe University”) that contain links to TargetDocuments of Concept “Homepage” and “Doe”. The links retrieved are of type “User Homepage” and are presented in an XML document created by the construct command.

The basic query categories that can be addressed regarding hyperlink information, along with some example questions and the relevant XML-QL queries follow.

- Queries on the destination or the source or the link type only:

Which pages point to “http://www.aueb.gr”?

```
    where <Document*>
    <SourceDocument URL=$U></>
    <Link><TargetDocument>
    <TargetObject      URL=”http://www.aueb.gr*”>
  </></>
    </></> in  www.a.b.c/hyperlinks.xml
 construct <link>$U</link>
```

- Queries on two or more of the source link and destination

Which university pages point to publications of AUEB?

```
    where <Document*>
    <SourceDocument URL=$U></>
    <Link><TargetDocument>
    <TargetObject URL=”http://www.aueb.gr*”> </>
    <LinkType Weight=$W Name=”Publication”> </>
    </></> in  www.a.b.c/hyperlinks.xml
 construct <link><URL>$U</URL>
              <Weight>$W</Weight></link>
```

- Queries that aggregate on the results

How many pages point to my biography page. (group by source subject)?

- Queries based on link type hierarchy

From the bibliography page select the publications (books, articles) on XML.

- Weight based sorting.

Return a list of links to my paper, that present contradictory or supportive opinions. Order by relevance.

- Find cliques, hubs and authorities.

Return a list of authority pages concerning sports.

Since the user defines an ontology space, the link type and the concept hierarchy are known to the query engine. A choice of concepts, instances, or combinations of both may be issued to the user in *HTML forms*. Choice options may be selected through check boxes, selection-lists or radio-buttons, thus providing an efficient user interface.

5. Conclusions

Web information systems face many challenges. Some of them follow:

- Lack of explicit representation and manipulation of links' semantics. Existing approaches are limited in providing a text value to describe the links contents. Additionally the uncertainty of the logical relation is not considered.
- Hypermedia network is barely supported. An example is the incomplete update of the system in the case of page insertion or deletion.
- Lack of homogenous representation of hyper-documents as well as of media objects and applications that are inserted to them.

Previous research efforts on hyperlinks address several issues that must be taken under consideration when creating, implementing and validating an information model. As stated by Kopetzky [13] typed links suffer from insufficiency, which poses some issues that are treated in the following.

i) Unless a commonly accepted ontology is given for a field, the semantics of a given link type are known to the user but not to the reader. The latter must infer these semantics either from the type mnemonics from formal or informal descriptions.

- The model presented in this paper is generic and various user communities can adopt and use it on their own ontologies. Modified XML-Schemas that will support any possible concept hierarchy can be created easily.

ii) Even if both author and reader have a common understanding of link types, the type information is just one kind of abstraction of the target contents.

- This abstraction, although it is ambiguous can limit down the amount of data transferred to the user until he reaches his goal. This can speed up search and retrieval tasks.

iii) Additional authoring effort is required, so that typed links are neither a remedy for the bulk of existing WWW documents nor can they be expected to be soon applied by the myriad of Web designers active in the Internet.

- In a smaller scale and with the use of automatic tools the link information can be embedded faster. The manual process can be accelerated by the development of smart modifiable wrappers and interfaces that will be adapted to various sources.

An advantage of the current approach is that it supports a hierarchy of concepts that can be employed to educe relevance rules between web pages concepts. The hierarchy network poses certain limitations on the possible relations between two pages, thus defining a concrete frame for the definition of a knowledge network.

In our approach we create an enriched with knowledge web information system by emphasizing on link semantics (mainly on the logical type of a link and the rating of such classification). Further work will focus on integration of conceptual information in order to broaden the information retrieval scope. The implementation and integration of all the proposed components will be the next step of our work. The resulting system will be evaluated on real cases, using existing structured or semi-structured web data.

APPENDIX – XML-Schema Fragment

```
<?xml version="1.0"?>
<!DOCTYPE schema SYSTEM "xml-schema.dtd"
[<!ATTLIST schema xmlns:hyper CDATA #IMPLIED>]>
<schema xmlns="http://www.w3.org/1999/XMLSchema">
  <element name="LinkSet">
    <annotation>  <info> A LinkSet contains zero or more
inks
   from and to certain documents</info>  </annotation>
    <type>
      <element ref="hyper:SourceDocument"/>
      <element ref="hyper:Link" minOccurs="0"
maxOccurs="*"/>
    </type>
  </element>
  <element name="Link">
    <type>
    <group order="seq">
      <group order="choice">
        <element ref="hyper:SimpleSourceObject"
minOccurs="1" maxOccurs="*"/>
        <element ref="hyper:AdvancedSourceObject"/>
      </group>
      <element ref="hyper:LinkType" minOccurs="0"
maxOccurs="*"/>
      <element ref="hyper:TargetDocument"/>
    </group>
    <attribute name="LinkID" type="decimal"
minOccurs="1"/>
    </type>
  </element>
<element name="SimpleSourceObject" >
  <type>
    <group order="choice">
      <element name="Text"/>
      <element name="Image"/>
    </group>
  </type>
  </element>
  <element name="AdvancedSourceObject" >
    <type>
    <attribute name="Type" type="string" minOccurs="1">
      <datatype source="NMTOKEN">
        <enumeration value="Form"/>
        <enumeration value="EmbObject"/>
      </datatype>
    </attribute>
    </type>
  </element>
  <element name="SourceDocument">
    <type>
    <group order="seq">
      <element name="Concept" minOccurs="1"
maxOccurs="*"/>
    </group>
    <attribute name="URL" type="string" minOccurs="1"/>
    <attribute name="Extension"
type="hyper:DocExtension"/>
    </type>
  </element>
  <element name="LinkType">
    <type>
    <attribute name="Name" type="hyper:LinkTypeName"
minOccurs="1"/>
    <attribute name="Weight" type="hyper:Percent"
minOccurs="1"/>
    </type>
  </element>
  <element name="TargetDocument">
    <type>
    <group order="seq" minOccurs="1" maxOccurs="*">
```

```
      <element name="TargetObject"/>
      <element name="TargetFrame"/>
      <element name="Concept" minOccurs="1"
maxOccurs="*"/>
    </group>
    </type>
  </element>
  <element name="Concept">
    <type>
    <attribute name="Name" type="hyper:ConcepType"
minOccurs="1"/>
    <attribute name="Weight" type="hyper:Percent"
minOccurs="1"/>
    </type>
  </element>
  <element name="TargetObject">
    <type>
    <attribute name="URL" type="string" minOccurs="1"/>
    </type>
  </element>

  <datatype name="Percent" source="integer">
    <minInclusive value="0"/>
    <maxInclusive value="100"/>
  </datatype>
  <datatype name="LinkTypeName" source="string">
  <enumeration value="Citation"/>
      ….
  <enumeration value="Argument"/>
    <enumeration value="Solution"/>
</datatype>
  <datatype name="DocExtension" source="NMTOKEN">
    <enumeration value="html"/>
    <enumeration value="xml"/>
    <enumeration value="php3"/>
    <enumeration value="asp"/>
  </datatype>
  <datatype name="ConceptType" source="NMTOKEN">
    <enumeration value="typeA"/>
    <enumeration value="typeB"/>
    <enumeration value="typeC"/>
  </datatype>
</schema>
```

6. References

[1] M. Montebello, Optimizing recall/precision scores in IR over the WWW, ACM SIGIR 1998.

[2] Ronny Lempel, Shlomo Moran, The Stochastic Approach for Link-Structure Analysis (SALSA) and the TKC Effect, WWW9.

[3] C. Petrou, D. Martakos, S. Hadjiefthymiades. Adding semantics to hypermedia towards Link's enhancement and Dynaming Linking. HIM 1997.

[4] John Kleinberg. Authoritative sources in a hyperlinked environment Proc. ACM-SIAM Symposium on Discrete Algorithms 1998.

[5] Ellen Spertus. ParaSite: Mining structural information on the Web. Proc. 6th International World Wide Web Conference 1997.

[6] Y. Mizuuchi, K. Tajima, Finding Context Paths for Web Pages, ACM Hypertext 1999, Darmstadt Germany.

[7] S. Loh, L.K. Wives, J.P.de Oliveira, Concept-Based Knowledge Discovery in Texts Extracted from the Web, ACM SIGKDD Explorations, 2000.

[8] A. Broder, S. Glassman, M. Masasse, G. Zweig. Syntactic Clustering of the Web. Proc. 6th International World Wide Web Conference. 1997.

[9] James Allan. Automatic Hypertext Link Typing. ACM Hypertext '96, Washington DC USA. 1996.

[10] Randal Trigg. A network-based approach to text handling for the online scientific community. PhD thesis. University of Maryland, 1983.

[11] World Wide Web Consortium, HTML 4.0 specification, Links. (http://www.w3.org/TR/REC-html40/struct/links.html)

[12] Gene Golovchinsky. What the query told the link: The integration of Hypertext and Information Retrieval. ACM Hypertext '97. 1997.

[13] T. Kopetzky, M. Muhlhauser, Visual Preview for Link Traversal on the WWW, *8th International World Wide Web Conference*, Toronto, Canada, 1999.

[14] T. Phelps, R. Wilensky, Robust Hyperlinks cost just five words each, UCB Computer Science Technical Report UCB//CSD-00-1091, 2000. (http://www.cs.berkley.edu/~wilensky/robust-hyperlinks.html)

[15] R. Kreutz, B. Euler, K Spitzer, No longer lost in WWW-based Hyperspaces, ACM Hypertext 1999, Darmstadt Germany.

[16] T. Davenport, L. Prusak. Working Knowledge: How organizations manage what they know. Harvard Business School Press. 1998.

[17] S. Staab, J. Angele, S. Decker, M. Erdmann, A. Hotho, A. Maedche, H.-P. Schnurr, R.Studer, Y.Sure, Semantic Community Web Portals, 9th International World Wide Web Conference.

[18] Mark Berstein. Patters of Hypertext. ACM Hypertext '98. Pittsburgh PA USA. 1998.

[19] World Wide Web Consortium, XML Linking Language(XLink), Working draft 21-2-2000, (http://www.w3.org/TR/xlink/)

[20] A. Bonifati, S. Ceri, Comparative Analysis of Five XML Query Languages. (http://www.toriisoft. com/tech-papers/xmlsurvey.pdf).

[21] Jelliffe Rick, Document Type Definitions. A Technical Introduction, 1999. (http://www.ascc.net/xml/en/utf-8/seminars/pdf/DTD.PDF).

[22] G. Mecca, P. Merialdo, P. Atzeni..ARANEUS in the Era of XML, IEEE Data Engineering Bullettin, Special Issue on XML, September, 1999.

[23] N. Sundaresan, J. Yi, A. Huang, Using metadata to enhance a web information gathering system. Web DB 2000.

[24] C. Jenkins, M. Jackson, P. Burden, J. Wallis, Automatic RDF Metadata Generation for Resource Discovery, *8th International World Wide Web Conference*, Toronto, Canada, 1999.

[25] M. Fernandez, J. Simeon, P. Wadler, An Algebra for XML Query, 2000. (http://www.cs.bell-labs.com/~wadler/topics/xml.html)

[26] World Wide Web Consortium, XML-QL: A Query Language for XML.1998 (http://www.w3.org/TR/NOTE-xml-ql)

[27] Donald D. Chamberlin, Jonathan Robie, Daniela Florescu: Quilt: An XML Query Language for Heterogeneous Data Sources. WebDB 2000

[28] V.Christophides, S.Cluet, J.Siméon, On Wrapping Query Languages and Efficient XML Integration. ACM SIGMOD Conference 2000, Dallas, Texas, USA.