A Semantic Similarity Approach to Electronic Document Modeling and Integration

William W. Song*, David Cheung, and CJ Tan
E-Business Technology Institute, The University of Hong Kong
Hong Kong, PRC

Email: {wsong, dcheung, ctan}@eti.hku.hk

Abstract

The World Wide Web is an enormous collection of information resources serving for various purposes. However the diversity of the Web information as well as the related formats makes it very difficult for users to efficiently search and obtain the information they require. The reason for the difficulty is because most of the information uploaded to the Web is unstructured or less structured. Many metadata models are proposed to response to this problem. These metadata models attempt to provide a certain kind of general description for the Web information to improve its structuredness. Although these documents consist in a largest portion of the Web information or Web resources, few metadata models are dealing with the ill-structured Web documents through analyzing their semantic relations with each other. In this paper we consider this large portion of the Web information, called electronic documents. We propose a metadata model, called EDM (Electronic Document Metadata Model). Using the metadata model we can extract semantic characteristics from electronic documents and then use the characteristics to form a semantic electronic document model. This model, inversely, provides a basis for analysis of semantic similarity between electronic documents and for the electronic document integration. The document modeling and integration will support further manipulations on the electronic documents, such as exchange, search, and evolution.

1 Introduction

1.1 Problems

It. is known to all that the largest part of the Web resources consists of electronic documents. Electronic documents are also the major targets for search. However, it is an extremely difficult to effectively find one single electronic document that a user asks for out from an explosively large number of the Web documents. This difficulty arises because there is no well-defined structure to represent these documents. Although a Web document is designed and uploaded to the Web to convey certain information to the Web readers, the Web document *per se* does provide little information effectively for purpose of search. In other words, a Web document cannot be directly used to uniquely identify the Web document to be searched for.

Most of the search engines available use keywords (textual strings) matching mechanism to search the Web documents and other resources as well. This keyword match method implies that the Web documents must contain a list of presumed keywords. Because of subjectivity of assigning keywords and multivocality of keywords selected, search results by keywords are not satisfactory, for example, low in search precision and difficult in comprehension of the search results [7].

The introduction of labels (or ratings) [15] to describe a Web document was considered to be a better way to identify Web documents. The labels assigned to Web documents were pre-defined, to some extent, with a fixed meaning. More important is that these labels were related to each other according to a semantic model. For example, the RSAC*i* rating standard for recreation materials suggested by Recreation Software Advisory Council [17] provides a tree-like semantic structure for labels and ratings. In this semantic model, the Web documents labeled by a same label indicate that they belong to the same class. In other words, the documents in the same class have the same label. Motivation of deploying labeling techniques is to group Web documents in classes. Significant relationships between the classes are also defined according to the semantic data model adopted.

Consequently, by using the Web browsers having capability of rating the Web documents, the Web users can rapidly access to or ignore the Web documents in the certain class identified by the label [6]. Furthermore, the labeling system (i.e. the semantic model) may provide classifications for labels as well. We should note that such classifications or labeling methods can be defined by the Web resource providers or some authorities (e.g. a rating bureau) or both.

However, at least two problems occur in the labeling methods. First, a labeling method is not really a well-defined and theoretically sound semantic model. It is only an empirical model. Second, many Web-authoring tools do not provide the document authors any support in document labeling.

1.2 Conceptual modeling

Grouping electronic documents according to some prerequisite criteria is an important issue, which has recently received increasing attention from both the researchers and users. Its aim is to collect electronic documents related to a certain topic for a group of people having similar interest. These people sharing the same interest in certain subject form a newsgroup [1]. To find out the resemblance among these electronic documents on the topic and compute their similarities is a crucial step in grouping the electronic documents. The similarity

[*] Correspondent author.

0-7695-0577-5/00 $10.00 © 2000 IEEE

computation is based on the characteristics or attributes - we assume that they can be captured from the information or data hidden in the electronic documents.

Through observing keywords and labels (considered to be metadata) from the electronic documents, we found it quite important to suggest a description framework, which can capture as many as possible characteristics or metadata about electronic documents. In other words, we need a conceptual model for description of metadata, or metadata model as quite often called in publications, to support modeling and managing the electronic documents. Such a metadata model would, can define attributes for the electronic documents and relationships between the electronic documents, and hence assist the electronic document similarity computation [9]. The techniques of using the characteristics of conceptual objects for computing object similarities and then integrating them to form a new conceptual object have been well developed in the conceptual database design area [10]. Therefore, in this paper, we will use these similarity techniques for electronic document clustering.

We have noticed that there were many efforts put in improving various search mechanisms and techniques in order to improve search quality. Such search quality, such as precision and comprehension [7], is used to evaluate search engines about their search results [3, 8]. However, little work has been done aiming at analyzing electronic documents, extracting their characteristics (metadata formula), and hence defining a conceptual metadata model for describing and modeling the electronic documents. A conceptual metadata model is fundamental and essential for making full use of the electronic resources for the following reasons:

1. A conceptual metadata model captures the basic features of electronic documents than just superficial observations.
2. Conceptual modeling will organize these attributes on documents and relationships between document for the electronic document integration.
3. Modeling process will help us to have an abstractive view of the electronic document structure and a detailed description of the document characteristics.
4. More importantly, more than a decade's research and development on conceptual modeling has formed a sound basis for metadata research and application.

1.3 Related work

Many efforts have been put in various aspects of the electronic information applications, such as electronic document modeling, document similarity computation, electronic information search quality study, etc. The electronic document modeling aims to extract characteristics, or attributes from existing electronic documents and therefore to form a metadata model. Conversely, the metadata model can then support to formally define inter-document relationships and the relationships within the electronic documents.

A number of metadata models have been proposed and some of them have been recommendations of W3C. Among others are XML Schema (extensible markup language) [18], RDF (Resource Description framework) [16], and MCF, (meta-content framework [14]. These proposed metadata models share a common set of features. Their aims include describing the structure of Web sites, distributing annotation and authoring, and exchanging formats of information. For example, RDF contains a set of directed labeled graphs consisting of a set of nodes, labeled arcs, and attributed values, corresponding respectively, for example, Web resources, the relationships among the resources, and the attributes to describe the resources. RDF can be viewed as a very general data model for description of electronic documents.

However, we maintain that a conceptual metadata model should first of all take into consideration electronic document structures. From the document structures a metadata model can be defined relatively general in order both to effectively represent the common features or attributes of electronic documents and to easily apply the data model for different purposes, such as document clustering computation.

In [13], a set of document structures is defined, including sequence structure, grid structure, tree structure, and *Web* structure. This description of the Web document structures is mainly based on how the documents are related to each other and follows the criteria of predictability, information richness, and modifiability. The first term, predictability, indicates that it is easy to find related resources and users would not be lost in a chain of search processes. Information richness requires that a Web resource be linked to many other resources to gain more information on a subject. The term modifiability means that changes on a Web document would not cause substantial loss of related information, i.e. links to other Web documents.

Electronic document resemblance computation is to group together the documents of a common interest into one class having one or several common characteristics. For example, news and articles about Intranet will be put together in a special interest group, e.g. IntranetSig. Some approaches to similarity computation [3, 8] have been suggested and basic process can be described as follows. 1) A profile of record for the user's interests is collected and organized in certain forms. 2) Pick up one profile as an original and compare it with other profiles, and weigh the similarity distance between the picked profile with the other profiles. The shorter the similarity distance, the more similar are the original and the profile from the others. 3) Given a fixed distance value, all the profiles having the similarity distances to the original less than the value are considered to have the same profile items. 4) This set of profiles will be used as identifying standard for electronic documents and recommended to the users.

Searching Web information can be seen as to find a set of Web information items with one or several common features by giving one or several searching attributes. There exist a number of commonly used search engines for the Web readers to find information they need. However, arguments take place from time to time on the features of search engines, such as imprecision and incomprehension, as well as the search strategies deployed by the search engines. There are also quite a number of search methods or tools sprung out, declared to be an improvement or enhancement of those commonly used search engines. The improvements can be simply summarized as: 1) adding more capability of searching than merely textual; 2) taking into account of hyperlinks which are viewed as just texts by ordinary search mechanisms; 3) considering information clustering (grouping Web documents) according to some predefined profiles [2, 4, 12].

1.4 Paper Structure

As we discussed previously, conceptual modeling is a key to the analysis and formation of the structure of the electronic documents and their relationships. The conceptual model for electronic documents can also be used for the management of electronic documents, such as searching and analysis. In the next section, we will propose and discuss an electronic document metadata model, called EDM, and its constructs, where two supporting conceptual models, Basic Metadata Model (BDM) and Path Tree Structure (PTS) are described. In section 3, based on the metadata model, WDM, we analyze various relations between electronic documents and define a set of semantic relatedness relations and a set of semantic similarity relations. In section 4, we suggest a process of electronic document clustering and discuss the major steps of the process. Finally, in section 5, we conclude the paper and suggest our future work.

2 Conceptual Metadata Modeling

Previously we have briefly discussed the necessity of introduction of a conceptual metadata model for the electronic document clustering or integration. Due to the diversity of electronic document descriptions, searching, exchanging, and management of electronic documents are difficult. The diversity also causes the difficulty in direct use of existing conceptual modeling methods for purpose of electronic document modeling. It is indispensable to build up a conceptual metadata model, which is able to take into account various characteristics of the electronic documents. So we propose a conceptual metadata model for the description of electronic and Web documents.

The metadata model, called EDM (electronic document metadata model), is a conceptual model. It is intended to formalize various characteristics from electronic documents and various relations (attributes) between electronic documents. The metadata data model is supposed to serve for a few purposes, including

- to build up better Web documentation languages for Web document authoring,
- to use the attributes for classification of electronic document schemas, generated from the Web data mode, and
- to define similarity relations between electronic documents for clustering Web information.

In the following, we first discuss what information or knowledge we observe on the electronic documents. We suggest a layer structure for description of the electronic documents. Then in section 2.2, we propose a basic metadata model, on which the rest of electronic document metadata can be built, and a path tree structure for URLs. In section 2.3 we discuss EDM, the conceptual metadata model for description of electronic documents. Multiple layer (tree-like) structure is often used in document management. So we also consider this kind of structure as part of the metadata model. Finally, we propose a metadata model, EDM, which is to support electronic document grouping, classification, filtering, and intelligent searching.

2.1 Meta-information on electronic documents

Look at an electronic document. It usually contains a sequence of textual paragraphs separated by a number of textual headings. Within the paragraphs spotted are some underlined words, called hyperlinks. The hyperlinks connect the words to other electronic documents, which are supposed to provide further detailed information about the words. In addition, the window containing the electronic documents may be split into 2 or 3 frames to show certain kind of content cohesion.

In the HEAD part, if the Web document was written in HTML, we will see some "meta" information items and other information items, such as "title" of the page "keywords" used for searching the page, etc. Some Web documents, like http://www.w3.org/PICS/, contain PICS labels in their head part. PICS (platform for Internet content selection) is a metadata model for rating Web documents in order to filter the Web documents. Whether or not we can use PICS labels for Web information filtering is depended on if the Web browsers support to interpret and execute a pre-defined label taxonomy structure. From these descriptions we can see that an electronic document, together with other information, not visible to the end users, provides meta-information for identification of the electronic document in one way or another. In the following we summarize our observations and put forward a four-layer structure as meta-information for electronic documents.

We maintain that there are *four types* of information (metadata), which can be used to describe an electronic document. These descriptive information collections form four layers for describing electronic documents or resources along different dimensions, from different views, and for different purposes. The first layer is the information describing the object content, called **Content Information**. For example, consider the Web page of CNN. The fist, major goal of the Web page is to convey a variety of news. The descriptive information serving this goal in general includes headlines, subjects, introductory paragraphs, together with images, movies, and so on so forth. Within an article, there are subtitles, section titles, keywords, review comments, etc. By browsing the descriptive information, the Web readers can quickly figure out what the content is all about.

The second layer contains the *data items* used to describe the relevant to an electronic document, such as author, creator, creation date, etc. We may say that most of the attributes defined in the metadata model Dublin Core [9] to some extent belong to this layer. We call them **Managerial Information**. One of the important attributes within the managerial information is version. Because versioning is now an important measure of the evolutionary process of an object, this attribute is a dynamic factor about the document. The managerial information is essential when we want to know some "publication" information about the objects of interest. Most of these items in the metadata information provide classification and categorization information for management of objects.

Referencing Information, the third layer, comes from the "hyperlinks" appearing in an electronic document. We extend "links" to a more general concept to represent "reference links" to any Web information, documents, and resources. So the environment information can be also called reference information, which means that there may be other objects or resources associated to the focused object and used for detailed descriptions of the object. This information also contains for example a structure of an electronic document. For example, a paper's structure is presented with a set of links to its various chapters, which appear in other Web resources. The referencing structure for an electronic document can be hierarchical, where an upper object has several links to its children object, and neighboring, like in an electronic map where a number of links from a spot to its four direction neighbors.

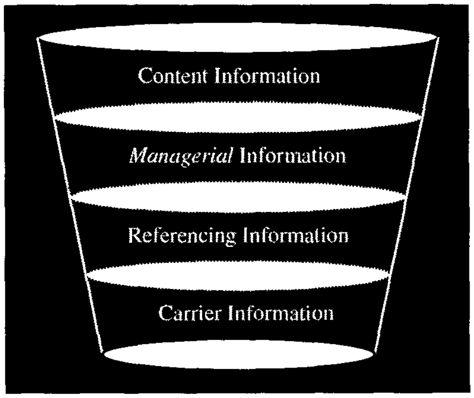


Fig. 1 Related information about an electronic document

The final layer is the **Carrier Information**, which in general provides the physical attributes about an electronic document. The information includes for example fonts, color, size, and so on. These information pieces become important when some semantic meanings are assigned to them. For example, "bold face" of a text may mean that the text is emphasized. In addition, in Email systems, people may hope to control the size of emails. Templates' information (layout) of electronic documents is also included in the carrier information, because this can help to manage different formats for documents.

2.2 Basic Metadata Model/Natural Tree Model

2.2.1 Basic Metadata Model

Through analysis of the electronic documents and the existing metadata modeling methods, there are some elements or element types in electronic documents. First, any document can be seen to be a resource. Second documents are linked together by some relationships, for example, the button *Next* in Powerpoint documents. Third, each document can be described a set of descriptive data, usually called metadata. Therefore, we can define a basic metadata model, which contains primitive constructs for electronic documents. The primitive constructs of this model are object, relationship, and attribute. Now we try to define the basic metadata model, denoted to be BDM.

Def 1 A basic metadata model is a triple, denoted as BDM=<O, R, A>, where O is a set of object types, R is a set of relationship types which relate one document to another, and A is a set of attribute types which describe a document.

Def 2 An instance of basic metadata model BDM is called a BDM schema, denoted as BDMS=<sn, o, r, a>, where sn is the unique name of the schema, o is a set of objects, r is a set of relationships, and a is a set of attributes.

The aim to define BDM schemas is to support us to analyze instance documents.

Def 3 A BDM object is a triple, denoted as BDM O=<on, or, oa>, where on is the name, or is the set of relationships from and to the object O, and oa is a set of attributes describing the object O.

Def 4 A BDM relationship is a triple, denoted to be BDM R=<rn, ro1, ro2>, where rn is the name, and ro1 and ro2 are BDM objects respectively.

Def 5 A BDM attribute is a triple, denoted as BDM A=<an, ao, av>, where an is the name, ao is the BDM object that A is to describe, and av is a set of values.

2.2.2 Natural Tree Structure Model

In information analysis and representation, information distribution structure is very important. Usually, people tend to organize documents according to certain criteria, for example, addressing the same subject or belonging to the same type. In organizing electronic pages, the "identifier" for a page is usually its URL (universal referencing location), giving a path (usually globally unique) to the page. Along the path, there may be more documents, each having its own URL but having the same domain name for example. In this sense, we consider a tree-like structure associated to such path-based URLs, more general, URI (universal reference identity). When we search for an electronic page, we may actually receive the other pages on the path to the page we require. More importantly these pages may support us with a better understanding of the meaning of the required page. In other words, the electronic documents on the same path or referencing link to the required document tend to have stronger semantic relations and tend to be clustered in one class.

For example, this paper, "A Semantic Similarity Approach to electronic Document Modeling and Integration", is found at the web site of http://www.eti-.hku.hk/pubs/meta-data/electronicdocuments/WebDoc-Model.html. We may reasonably assume that the other documents found with this directory of http://www.eti-.hku.hk/pubs/metadata/electronic-documents/ may deal with the similar problems related to metadata, electronic documents, etc. Here we deduce a group of documents having a closely related meaning or subject by the keywords or subjects from one of the documents.

Now we try to define this natural tree structure using BDM. It is obvious that the natural tree structure is a sub-model of BDM, because in the natural tree structure, although the objects can be any Web resources, the relationships between the objects follow the grouping semantics by the paths or the fragments of the paths to the Web objects.

Def 6 A path tree structure, denoted to be PTS=<O, R', A>, is a sub-model of BDM, where O and A are defined as the same as in BDM while R' is a subset of R. There is a particular object in PTS called root object. There is a set of objects in PTS called leaf objects.

Def 7 A branch relationship type R' in PTS is defined as R'=<rn, ro1, ro2>, where rn is the name cut from the path name, and ro1 and ro2 are respective a parent object and a child object.

2.3 WDM: A Conceptual Data Model for Electronic Documents

2.3.1 Metadata elements

In the section 2.1, we described the four layers of metadata information for electronic documents, i.e., content information, managerial information, referencing information, and carrier information. These four layers of metadata information consist in fundamental components in a conceptual metadata model, which is considered to be an extension to the basic metadata model, BDM. The four layers of metadata information describing the electronic documents can be either of relationship types between electronic documents or of attribute types describing electronic documents.

Now let us have a detail look at what are included within the four layers of metadata information.

Metadata on content (Content Information) include topic, title, subject, abstract, keyword, heading, sub-heading, content-type[1], etc. These data are considered to be attribute types.

Metadata on managerial elements (Managerial Information) include author, date, creator, version, edition, publisher (if any), number of pages, etc. The metadata are considered to be attribute types.

Metadata on referencing information include various links, connections, and other relationships. So these metadata will be mainly part of relationships in BDM.

Metadata on carriers (Carrier Information) include document types, e.g., Word, fonts, faces, media, etc. These are also considered to be attribute types.

In the reality, we divide the metadata information into two levels. The level-one metadata information about an object provides direct evidences of telling what the object is, while the level-two metadata only provides the information of telling what the object could be. For example, about the metadata on content, we may say that titles, subjects or keywords can be of level-one attributes, whereas content-type, sub-headings are of level-two attributes because their contribution to the semantic identity of the Web object[2] is not as significant as the metadata of level-one attributes.

Similarly, the two-level division is also valid to the metadata on managerial information and the metadata on carriers. In the reality, whether a metadata belongs to the first level or the second level is depended on the users' judgements on how important the role the attributes play in the object semantic identification. Therefore such division is of high subjectivity. This division is on the purpose of electronic document clustering and integration. It is true in the reality that some characteristics are more important than the other metadata in identifying objects.

2.3.2 Metadata Model: EDM

As we discussed previously, an electronic document contains many descriptive items, called metadata. Some metadata are more important and useful in identifying an electronic document than the other metadata. Therefore the former metadata are the level-one attributes and the latter the level-two attributes. In the following, we give formal definitions to EDM, the electronic document metadata model. First of all, we can say roughly that EDM is a sub-model of BDM, because the set of objects in EDM (only electronic documents) is a subset of the set of objects in BDM (any objects).

Def 8 (WDM) The metadata model WDM, denoted to be WDM=<DO, DR, DA>, is a sub-model of BDM, where DO is a subset of O, DR a subset of R, and DA of A.

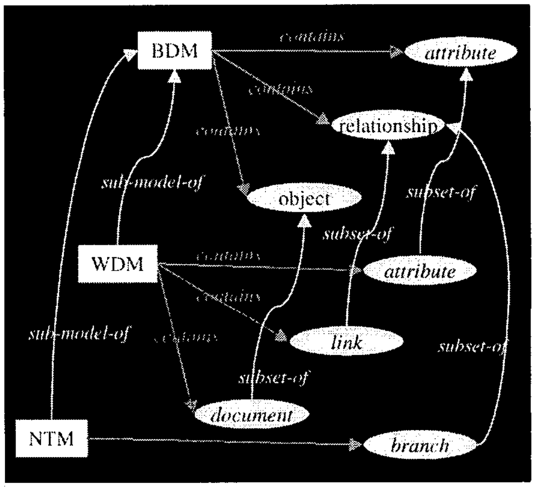


Fig. 2 Meta-model of WDM, PTS, and BDM.

In the context of the Web application, DO is a set of electronic documents, DR a set of relationships linking one electronic document to another, and DA a set of attributes describing electronic documents.

Def 9 A WDM document is a triple, denoted as WDM DO=<dn, dr, da>, where dn is the document identifier, dr is a set of references from and to the document, and da is a set of attributes describing the document.

Def 10 A WDM reference is a triple, denoted to be WDM R=<rn, rd1, rd2>, where rn is the name, and rd1 and rd2 are BDM documents respectively.

Def 11 A WDM attribute is a triple, denoted as WDM A=<an, ad, av>, where an is the name, ad is the WDM document that A describes, and av is a set of values associated to the document ad through A.

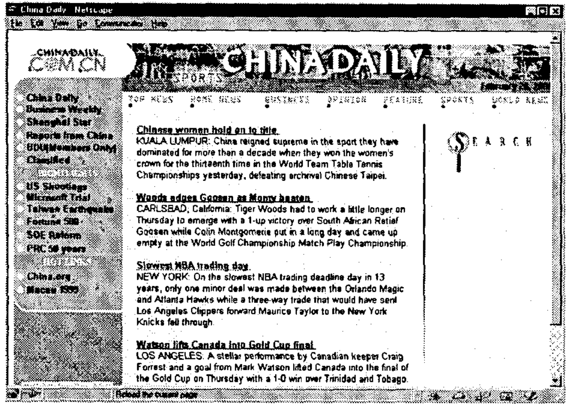
These definitions can be graphically illustrated by the meta-model (self-explanatory) as in the figure Fig.2.

[1] Content-type indicates that metadata tell what kind of narrative types is related to the document, such as report, memo, minutes, etc.

[2] When no obvious misunderstanding arises we interchangeably use the terms electronic document, electronic object, or electronic resource.

2.3.3 Example of WDM

In this section, we consider an example to illustrate the concepts proposed in the previous section. The example is a fragment of a Web document displaying news. Here we try to extract the metadata out from the document for describing the document.



Chinese women hold on to title.
KUALA LUMPUR: China reigned supreme in the sport they have dominated for more than a decade when they won the women's crown for the thirteenth time in the World Team Table Tennis Championships yesterday, defeating archrival Chinese Taipei.

Woods edges Goosen as Monty beaten.
CARLSBAD, California: Tiger Woods had to work a little longer on Thursday to emerge with a 1-up victory over South African Retief Goosen while Colin Montgomerie put in a long day and came up empty at the World Golf Championship Match Play Championship.

Slowest NBA trading day.
NEW YORK: On the slowest NBA trading deadline day in 13 years, only one minor deal was made between the Orlando Magic and Atlanta Hawks while a three-way trade that would have sent Los Angeles Clippers forward Maurice Taylor to the New York Knicks fell through.

Watson lifts Canada into Gold Cup final.
LOS ANGELES: A stellar performance by Canadian keeper Craig Forrest and a goal from Mark Watson lifted Canada into the final of the Gold Cup on Thursday with a 1-0 win over Trinidad and Tobago.

Fig. 3 A Web document – Web newspaper.

Fig.3 is a Web document, about the subject of Sport News. The document contains a number of references to its children documents. We assume that the document can be uniquely identified by its internal series number. The content of the document can be described by its super-title, its hierarchical table of contents, and its introductory paragraphs. That is, there are two attributes, <000226, doc_title, "China Daily"> and <000226, Tree-Struct-6, "Sport News">. In addition, the document owner name, logo, the referencing links are also important attributes. Now we use the four layers of metadata to classify some of the metadata information in Fig.3 and represented in the following statements:

Metadata on Content:
<000226-1, title, "Chinese women hold on to title">
<000226-1, leading_paragraph, "Kuala Lumpur: China reigned supreme in sport ...">

Metadata on Managerial information:
<000226-1, written_by, "Mr. Wong">
<000226-1, date, 2000-02-26>
<000226-1, on, "China Daily">

Metadata on Referencing information:
<refer_to, 000226-1, Sport_News>
<next_doc, 000226-1, 000226-2>

Metadata on Carrier information:
<000226-1, doc_type, text>
<000226-1, format_type, report_type>

2.3.4 Summary

We have formally introduced the electronic document metadata model, EDM, and some related concepts in the previous sections. Our main attempt is to base the electronic document clustering on a formal representation. This formal representation of EDM can support similarity comparisons and representations in section 3. For example, by comparing two attributes and their corresponding values of two documents, we can find out in which way they are related to each other, e.g., content similar or carrier similar. The use of finding out carrier similar of a number of electronic documents lies in the possibility that users can customize their metadata tools.

In our opinions, main requirements that should be met by a conceptual metadata model as EDM include 1) easy to understand and to use, 2) capable to represent other metadata models, and 3) well-defined for the electronic document analysis and classification. One of the reasons to propose these requirements is that the metadata model should be general and expressive enough to translate various object types and relationships existing in the electronic documents and hence easier to represent them for purposes of analysis and classification [11].

3 Semantic Relations and Similarities

The EDM, we described in last section, provides an important modeling basis for clustering electronic documents. In this section, we will discuss various relatedness and similarity relations between electronic documents. As we know, electronic documents can be related to each other in a variety of ways. Possibly two documents have exactly the same content, or their headings indicate that one document is a follower of another (like section 1 and section 2), or their links reveal a referencing relation (such as a detailed explanation of a phrase). These situations are considered to be a motivation for electronic document clustering or integration[3].

Our consideration on the electronic document relatedness as well as similarity relations are based on the assumption that two electronic documents are more strongly related (probably more similar) to each other if their components have more in common. In other words, two electronic documents are similar if their attributes and relationships are respectively similar. In order to cluster electronic documents together based on their relatedness relations we also assume that these schemas can be included in the same cluster if some attributes of WDM schemas are partially the same.

3.1 WDM Similarity Classes

Because of electronic documents being complicated, changeable, and less structured, WDM suggests a structured and formal means to modeling electronic documents. The metadata modeling support includes finding out characteristics from documents and inter-document relationships, comparing and classifying the characteristics on a formal basis, and grouping (clustering) the analyzed documents.

Based on the WDM model, we consider a classification of possible relations between the electronic documents to be compared. This classification is also based on the following three assumptions. The first assumption is that a pair of documents are related if any pair of their attributes or relationships are related. The second assumption is that the determination of inter-document semantic relations is upon the users' empirical and conceptual understanding of the words or phrases used in the significant attributes or relationships of the documents. For example, within the conceptual database area these two words, semantic and conceptual, are

[3] Sometime we consider that electronic document clustering and electronic document integration are synonyms.

considered to have the same meaning. Therefore, two documents can be considered to be closely related (or addressing similar subject), if one document title is "on semantic data modeling" and the other document title is "on conceptual modeling".

This assumption leads to a third assumption, which is only used for easy narrative. The third assumption is, when we say that two document components (attributes or relationships) are related, we mean that they have common words, phrases, etc. in key places. However, they don't have to be semantically related. For example, two documents have the same carrier, Word or CD-ROM. However, the restriction on the semantic similar relations of electronic documents will be much stronger. Two documents being semantically similar means they should address on a very close subject. How to make quantitative measurement on document semantic relations or even semantic similarities is still a question and will be addressed in our next step of metadata research.

3.2 Semantic Relatedness Relations

According to the assumptions discussed above, we define a set of relatedness relations between documents. These definitions include semantic relatedness relations on Content, Managerial, Referencing, and Carrier as we discussed in the section 2.1. In addition, we also define the relatedness relation on the Path or Branch relationships defined in the natural tree model, PTS. We believe that the Path relationship plays an important role in identifying an electronic document. In the following we will use content attributes, managerial attributes and carrier attributes to represent the attributes for the content information, the managerial information, and the carrier information of electronic documents respectively.

Def 12 (Content-relatedness) Two WDM documents are content-related if they at least one same content attribute.

Def 13 (Managerial-relatedness) Two WDM documents are considered to be managerial-related if they have at least one same managerial attribute.

Def 14 (Path-relatedness) Two WDM documents are considered to be path-related if their URLs have the same domain prefix.

Domain prefixes are the part, e.g. http://www.eti-.hku.hk/pubs/, of URLs. In the section 2.2 we have defined the natural tree model. This model is particularly used to describe and represent the documents identified by URLs. Here we can intuitively assume that the documents having the same URL prefix possess shared characteristics.

Def 15 (Referencing-relatedness) Two WDM documents are considered to be referencing-related if their referencing names are similar[4].

Def 16 (Carrier-relatedness) Two WDM documents are considered to be carrier-related if their carrier attributes are similar.

In addition, regarding the attributes to the content, managerial, and carrier information of electronic documents, we also consider their two levels of significance of the attributes contributing the document identification. Roughly speaking, the level-one attributes are more important in judging two electronic documents to be semantically similar than the level-two attributes.

In the following table, Table 1, we give a basic and subjective estimate of the various attributes contributing to the semantic similarity between electronic documents. Generally and intuitively, the content attributes give more information about the content of an electronic document than the other attributes. Similarly, the managerial attributes give more information about the content of an electronic document than the carrier attributes. On the other hand, the users may require grouping electronic documents in one cluster rather by some particular attribute values, e.g. by the same publisher, than the meaning closeness of the documents, e.g. under the same subject.

	Content	**Managerial**	**Carrier**
Level-1	**5**	**4-5**	**2-3**
Level-2	**4-5**	**3-4**	**1-2**

Table 1 Basic estimate of similarity contribution.

However, in order to provide a quantitative measure of semantic similarities between electronic documents, we need a set of figures. Furthermore, since we consider a two-level division of the attributes, the basic estimate of similarities would support the accuracy of the similarity comparison and analysis. In the table, we use a scale of 1 to 5 to measure the significance of one type of the attribute compared to the other. 1 indicates the attribute contributing least to the similarity comparison and 5 the most. In the next section, we will use this estimate for the semantic similarity definitions.

3.3 Semantic Similarity Relations

Once a relatedness relation is found between two electronic documents, we begin to consider whether the pair of electronic documents is semantically similar. In other words, we start to consider the semantic similarity relations between electronic documents. Semantic similarities are the basis for grouping or clustering electronic documents together. The previous definitions of the relatedness relations and the basic estimates of the various attributes contributing to the semantic similarity measure provide us with a step toward the semantic similarity definition. It is obvious that all the semantic similarity relations between electronic documents are a subset of the relatedness relations.

In the following we define four kinds of similarity relations.

Def 16 (Content-similar) Let d1 and d2 be two WDM documents, a1 and a2 be two content attributes of d1 and d2 respectively. The documents d1 and d2 are content-similar if a1 and a2 have the same name and the same value.

Def 17 (Managerial-similar) Let d1 and d2 be two WDM documents and a1 and a2 be two level-one managerial attributes of d1 and d2 respectively. The documents d1 and d2 are considered to be managerial-

[4] Here we borrowed a similar definition from [9] for the term semantic similar for names. Two names (or words) are semantic similar if they are same, synonymous, or close in meaning.

similar if a1 and a2 have the same name and the same value.

Def 18 (Carrier-similar) Let d1 and d2 be two WDM documents and a1 and a2 be two level-one carrier attributes of d1 and d2 respectively. The documents d1 and d2 are considered to be carrier-similar if a1 and a2 have the same name and the same value.

Def 19 (Path-similar) Let d1 and d2 be two **path related** WDM documents. The documents d1 and d2 are considered to be path-similar if d1 and d2 share the same directory.

3.4 Electronic Document Integration

The establishment of the similarity relations between documents requires a series of comparisons between the corresponding elements or components of the electronic documents to be compared. The comparison results of one single pair of document elements may not show sufficient evidence to integrate the compared documents, but the comparison results from a significant amount of pairs of document elements will to some extent display the semantic relations between the compared documents.

An individual managerial attribute contributes to the identification of an electronic document not as much as an individual content attribute does, but a set of managerial attributes can be much more significant in identifying the document. Moreover, when the users expect to focus on e.g. an author of a number of papers, the managerial attributes will play a more important role than the other attributes in integrating the documents.

We consider that a major task in integrating documents is to measure the similarity relations among the electronic documents. In general, similarity measure method should be based on the qualitative, as well as quantitative analysis of the documents. Sometime the quantitative analysis of the documents is even more important because only good quantitative measure of the electronic documents' attributes make it possible to automate the process of electronic document integration. Due to the paper size, we will not discuss the quantitative analysis of the document semantic similarities in detail.

4 Document Integration Process

In this section, we propose a general architecture for the electronic document integration. To apply EDM and its semantic similarity relations to electronic document integration is an important activity in the architecture. Based on the description of the inter-document relations that we discussed previously, we consider that the process from electronic document modeling to similarity relation computation consists of six steps, see the figure below.

1) **Modeling electronic documents**. Use EDM, to describe electronic documents and extract their components, such as content, managerial, and carrier attributes, as well as various relationships between the documents. Then construct the WDM schemas based on the description and extraction of the electronic documents. This step is usually called conceptual modeling.

2) **Sorting out the document components**. The description and extraction of the electronic documents is then stored in the repository for the semantic analysis in the next step. This storage of the electronic documents contains two indexes. One is the index of the WDM schemas and the other is the index of the documents. The schema index is useful for relationship comparisons between the documents.

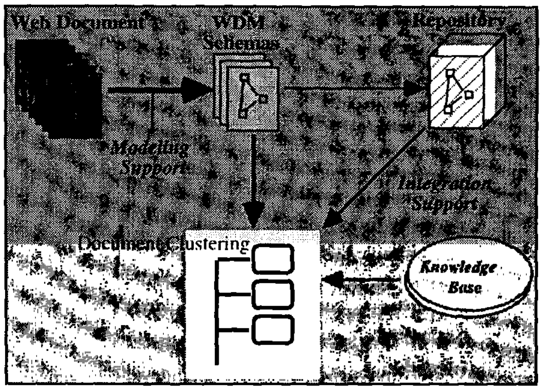


Fig. 4 A process of electronic document modeling and integration for intelligent searching

3) **Generating the relatedness relations**. A list of inter-document relatedness relations is generated by comparing in pair the WDM documents as well as the WDM schemas in terms of their attributes and referencing relationships in the repository (termed as vocabulary in some literatures). The result of this comparison, maintained in the repository, is a list of the document pairs, each of which contains two documents being related. Of course, one document may appear in more than one pair of electronic documents.

4) **Generating the similarity relations.** A list of inter-document similarity relations is generated from the list of the document pairs having the relatedness relations based on the semantic similarity method discussed previously. The similarity comparison result is also maintained in the repository. The result is a list of the document pairs, each of which contains two documents being semantically similar. Like in the relatedness relation pairs, it is allowed to one document appearing in two or more similarity comparison pairs.

5) **Quantitatively measuring the document similarities**. As the list of the document pairs of similarity is obtained, we begin the step of quantitative analysis to the documents. Based on the significance of attributes and relationships contributing to the document semantic similarity comparison, a scale of weights is given to the attributes and relationships. Then by using the weights, a group of quantitative distances is achieved to represent quantitative similarity relations between the documents.

6) **Establishing the document clusters**. Through the similarity analysis and quantitative computation, we will find some documents more closely related to each other. In other words, the semantic distances between these documents are much less than the semantic distances to the other documents or less than a threshold we predefined. These documents will form a cluster with many shared characteristics. Similarly, the other

documents may form one or more clusters. We admit, it is quite possible that a document belongs to no group when it is not closely related to other documents.

In the architecture, we also consider a knowledge base for semantic relation analysis. The knowledge base contains a set of rules, such as

same(a, b) and ***same (b, c)*** implying ***same(b, c),***

and a set of semantic definitions for concepts, e.g.

synonym(title, topic).

Here, the predicate ***same(a, b)*** means that a is the same as b, and ***synonym(title, topic)*** means that title is synonymous to b. The importance of maintaining this knowledge base lies also in knowledge accumulation, such as the knowledge of metadata models, for future use.

5 Conclusion

In the management and use of electronic documents, which consists in the largest portion of the Web information resources, two aspects are crucial. One aspect is how to describe and model these electronic documents. The other aspect is how to analyze and represent the electronic documents. In this paper, we have proposed a preliminary electronic document metadata model, EDM, for describing, structuring and modeling the electronic documents and defining a set of components for the documents. These components or elements of electronic documents are very useful for the document analysis and the inter-document relation computation. Based on EDM, we have also suggested an analysis method for determining the relations between electronic documents, i.e., relatedness relations and similarity relations. We structured a general process from the electronic document modeling to the document clustering.

However, as we have already seen, the determination of the document attributes is still a problem, in particular, the determination of the attributes, which contribute significantly to the determination of the inter-document relations. A second question is the determination of carrier attributes. It is gradually recognized that the carrier attributes are playing important roles in determination of document content, in particular, when many multimedia resources are available in the electronic.

Another question is, although the relationships between documents are critical in determining the document contents, how to break down the inter-document relationships, mainly the hyperlinks, into a set predefined classes. We also realized that the electronic documents do not exist in a knowledge-vacant space. Inversely, the Web is full of information and information on knowledge. This knowledge should be extracted for our document modeling purpose. We consider the path-relationships of electronic documents are a source for the Web knowledge discovery [5].

Our next step will be to refine EDM to improve the expressiveness of the model. We will also define a set of quantitative measurements for the inter-WDM document similarity computation so that the document clustering becomes more useful and practical both in the Web information search and in the electronic document management.

References

[1] Chang, C., et c. ***Customizable Multi-Engine Search Tool with Clustering.*** The 6th International World Wide Web Conference. Santa Clara, USA. 1997.

[2] LaMacchia, B. ***The Internet Fish Construction Kit.*** The 6th International World Wide Web Conference. Santa Clara, USA. 1997.

[3] Maltz, D. and K. Ehrlich. ***Pointing The Way: Active Collaborative Filtering.*** CHI'95. 1995.

[4] Marchiori, M. ***The Quest for Correct Information on the Web: Hyper Search Engines.*** The 6th International World Wide Web Conference. Santa Clara, USA. 1997.

[5] Ng, Chi-Yuen, Ben Kao and David Cheung. Text-Source Discovery and GLOSS Update in a Dynamic Web. The Fourth Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD-00), Kyoto, April, 2000.

[6] Resnick, P. and J. Miller. ***PICS: Internet Access Controls without Censorship.*** CACM. ***39(10).*** 1996.

[7] Shakes, J., et c. ***Dynamic Reference Sifting: A Case Study in the Homepage Domain.*** The 6th International World Wide Web Conference. Santa Clara, USA. 1997.

[8] Shardanand, U. and P. Maes. ***Social Information Filtering: Algorithms for Automating "Word of Mouth".*** CHI'95. 1995.

[9] Dublin Core, http://purl.oclc.org/dc/, 1999.

[10] Song, W., P. Johannesson and J. Bubenko. ***Semantic Similarity Relations and Computations in Schema Integration.*** Journal of Data and Knowledge Engineering. ***19(1).*** 1996.

[11] Song, W. ***WDM: A Web Document Model and Its Supporting Web Document Analyzer***, AIS'98, Baltimore, MD, USA, 1998.

[12] Spertus, E. ***ParaSite: Mining Structural Information on the Web.*** The 6th International World Wide Web Conference. Santa Clara, USA. 1997.

[13] White, B. ***Web Document Engineering***, Tutorial, The 5th International World Wide Web Conference, Paris, France. 1995.

[14] MCF, http://www.w3.org/Member/9706/xmlmcf.htm.

[15] PICS, http://www.w3.org/PICS/.

[16] RDF, http://www.w3.org/RDF/.

[17] RSACi, http://www.rsac.org/homepage.asp.

[18] XML, http://www.w3.org/XML/.