

Semantic Coordination for Document Retrieval

Bernardo Magnini, Luciano Serafini, Manuela Speranza

We present CtxMatch, an algorithm that finds mappings between two heterogeneous partially overlapping Classification Hierarchies (e.g. taxonomic structures used to organize documents). CtxMatch relies on the semantic interpretation of both the labels provided in the CHs and the hierarchical structures of the CHs; it does not consider the content of classified documents, thus allowing the retrieval of any kind of documents (e.g. text files, images, applications, videos, etc.). The Web Directories of Google and Yahoo! have been chosen as an evaluation set for discussing the performance of CtxMatch.

1 Introduction

Distributed approaches to Knowledge Management are based on the recognition that different communities in a complex organization (e.g. corporations, the Semantic Web, etc.) are autonomous and have their own partial and subjective conceptualization of the world. They organize their knowledge according to local schemas, which are used to classify, share, and update documents. Examples of schema models include ER-schema automata, local ontologies, taxonomies and Web Directories. Since different communities (and even people in the same community) may classify similar documents in different ways, the problem of interoperability between schema models (i.e. semantic coordination) becomes crucial for any document retrieval approach.

In this paper we focus on Classification Hierarchies (CHs), i.e. taxonomic structures used to organize large amounts of documents. The most typical examples of CHs are file systems, marketplace catalogs, and the directories of Web portals. Documents of a CH can be of many different types, depending on the characteristics and uses of the hierarchy itself. In file systems, documents can be any kind of file (e.g. text files, images, applications, etc.); in the directories of Web portals, documents are pointers to Web pages, while marketplace catalogs organize either product cards or service titles.

CHs are now widespread as knowledge repositories and the problem of their integration is acquiring a high relevance from a scientific and commercial perspective. A typical application of CH interoperability occurs when a set of companies want to exchange products without sharing a common product catalog. In this case the best solution is to find mappings between the catalogs (Schulten et al., 2001).

CHs are used for document classification and retrieval. Users browse hierarchies of concepts and quickly access the documents associated with the different concepts. The content of a concept is typically described by a label, but it also depends on the concepts at higher levels in the hierarchy. The relations between concepts in a CH are usually not explicitly labeled and their interpretation is ambiguous and context dependent. The arc connecting *Sports* and *Organizations* in the CH in Figure 1, for instance, is a specification arc, while the relation between *Clubs and Schools* and *United States* is a location relation.

The procedure commonly followed by users to retrieve documents in CHs is based on a semantic interpretation of the labels associated with the nodes, so that in most cases users do not need to check the content of the documents. Users enter the hierarchy from the root node, and, at each node, choose the child node under which the documents are more likely to be classified.

Consider, for instance, the CH depicted in Figure 1. In order to find documents about Romanian artistic gymnasts, a user would start from the root *Sports*, would first select *Gymnastics*, then *Artistic*, and then *Gymnasts*, and would finally retrieve the documents. Users' choices are guided by the following facts:

- Understanding of the meaning of the labels attached to the nodes encountered during the navigation.
- Knowledge of the fact that Romanian gymnasts are artistic gymnasts and that gymnastics is a sport.
- Assumption that a document about artistic gymnasts is more likely to be classified under the sub-tree rooted at *Sports/Gymnastics* than under the ones rooted at *Sports/Billiards* and *Sports/Organizations*. Similar assumptions are related to the choice between the children of *Sports/Gymnastics*, and so on.
- Awareness that the node *Gymnasts* is the most specific node about the topic 'Romanian gymnasts', as there is no node *Romania* available under *Gymnasts*.

These facts reflect the criteria in which documents are typically classified in CHs, such as those adopted in Yahoo! or in the Open Directory Project.¹

Semantic coordination offers a powerful view of an IR scenario in which both the user's query and the document collection are organized as CHs. The source CH (i.e. the user's query) and the target CH (the document collection) provide contextual and structured information which is exploited in order to semantically drive the retrieval process. Such a process is per-

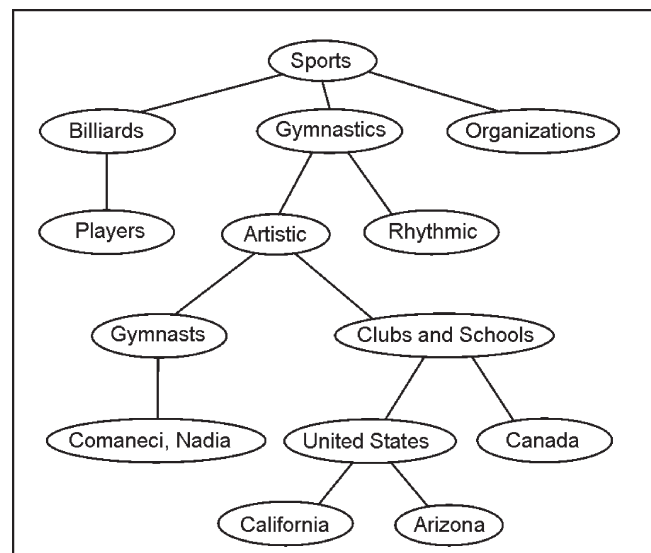


Figure 1: Example of CH (Google Web Directories).

formed by mapping the source node against all nodes in the target CH and selecting the nodes for which an equivalence (or inclusion) mapping to the source node holds.

The aim of this paper is to describe a method to analyze the implicit knowledge hidden in CHs and to make such information explicit in order to provide a correct interpretation of their concepts. In particular, we describe an algorithm that, given two CHs, returns an interpretation of each node in terms of a logical formula of description logic and computes the mapping relation between pairs of nodes. Unlike previous approaches to interoperability, we do not consider the content of the documents classified under each node, which allows us to work with any kind of document (e.g. text files, videos, images, etc.). Rather, we rely on the semantic interpretation of the labels describing the nodes, which is obtained through a linguistic analysis of the labels, and on the structure of the CH itself.

The paper is structured as follows. In Section 2 we review the relevant approaches to interoperability among CHs, outlining the main differences with respect to the semantic-based approach we propose. In Section 3 we introduce the architecture of the CTXMATCH algorithm and describe in detail the phases of the linguistic processing involved. In Section 4 we discuss the results of an evaluation experiment where CTXMATCH is applied to the Web Directories of Yahoo! and Google.

2 Approaches to CH Interoperability

In our view, the problem of the interoperability among different CHs can be roughly stated in this way: given a node N_s in a source CH and a node N_t in a target CH, a mapping algorithm has to discover a relation between N_s and N_t . Although there can be differences in the definition of the task itself (Agrawal and Srikant, 2001; Madhavan et al., 2002), and considering that this is a relatively new challenge, approaches to CH mapping can be grouped into four classes, according to the kind of information used: approaches which consider the content of the documents belonging to the CH; approaches based on the classification of the documents; approaches that take advantage of the structure of the CH; finally, approaches that attempt a semantic interpretation of the CH labels.

In the rest of this Section we will briefly review the first three approaches, while the semantic-based approach will be introduced in more detail in Section 3.

Mapping based on document content. These approaches rely on the content of the documents classified in a CH. As an example, the GLUE system (Doan et al., 2002) employs machine learning techniques to discover mappings among CHs. The idea consists of training a classifier using documents of the source CH, and then applying that classifier to documents of the target CH, and vice-versa. It has to be remarked that, at least for the experiment presented by the authors, the learner considers the *textual content* of the documents, i.e. they are managed as a bag of tokens. The major drawback of this approach is that it requires textual documents, which prevents its usability

when such documents are of a different nature (e.g. images) or they are not available at all.

Mapping based on document classifications. An improvement with respect to the content-based approach has been proposed by Ichise et al. (2003), who address the mapping problem computing a statistical model of the classification criteria of the source and the target CHs. Such a statistical model attempts to determine the degree of similarity between two categorization criteria considering the number of documents in common to nodes of different CHs. The advantage over the content-based approach is that the analysis of the documents is not necessary. However, it is required that the source and the target CHs have a certain amount of documents in common, a situation that in most of the concrete application scenarios is hard to obtain.

Mapping based on structural information. These approaches attempt to discover mappings independently of the number and the type of documents classified by the CHs. An interesting approach in this direction has been proposed by Daude et al. (2000), who exploit a constraint satisfaction algorithm (i.e. relaxation labeling) for discovering relations among ontologies. The algorithm first selects candidate pairs using lexical similarities (i.e. concepts with the same label) and then considers a number of structural constraints among nodes (e.g. the two nodes have their respective hypernyms connected) to increase or decrease the weights of the connection. Although the approach has been experimented and evaluated to map two different versions of WORDNET, achieving high accuracy, our impression is that mapping CHs is a sensibly harder task, due to the highly idiosyncratic way in which CHs may organize their content.

Among the class of structure-based approaches, we can also include work that specifically addresses the integration between ontologies expressed in a formal language (e.g. description logic). These approaches either aim at establishing the logical apparatus and properties for the task, e.g. (Calvanese et al., 2001), or propose semi-automatic procedures which need a final validation of the mapping from a domain expert, e.g. (Noy and Musen, 2001).

3 The CTXMATCH Algorithm

CTXMATCH is a particular implementation of an approach to *semantic coordination* that has been proposed recently (Bouquet et al., 2003). The algorithm strongly relies on a linguistic analysis of the labels contained in the CHs. The main difference between CTXMATCH and other approaches to schema matching (see Section 2) is that in order to interpret a node of a CH it considers the *implicit information* deriving from the *context* where the node occurs, i.e. the structural relations with the other nodes.

CTXMATCH takes as input two nodes in two different classification hierarchies and gives as output a mapping relation between them (see Figure 2). It consists of three main phases:

(i) *Linguistic analysis of the labels.* We analyze the labels attached to the nodes from a linguistic point of view. First, we perform shallow parsing using Alembic (Day and Vilain, 2000), then we perform a semantic interpretation of the labels exploiting the world knowledge contained in WORDNET (Fellbaum, 1998) and finally, we generate a formula in description logic (Baader and Nutt 2002) representing a first approximation of the meaning of the node.

1 Two fundamental criteria which are stated in many guidelines for Web Directories classification are the 'Get specific' criterion and the 'Look familiar' criterion.

Get Specific: When you add your document, get as specific as possible. Dig deep into the directory, looking for the appropriate sub-category. You can't submit your company to a top level category. [...] Dig deeper.

Look Familiar: Armed with the above knowledge, browse and search your way through the hierarchy looking for the appropriate category in which to add your company. Look for categories that list similar documents.

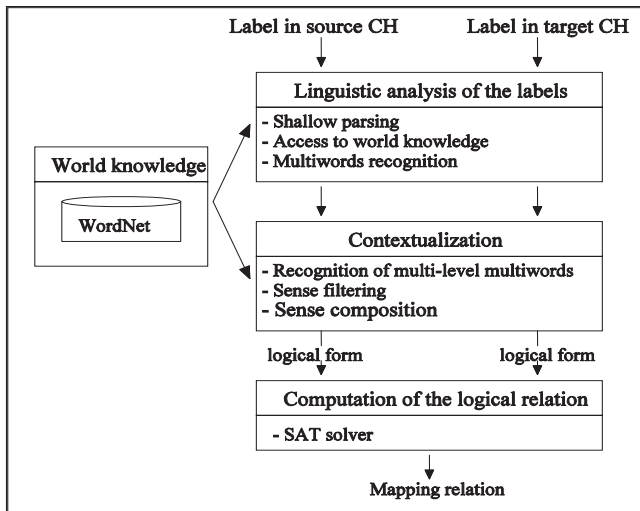


Figure 2: The architecture of CTXMATCH.

(ii) *Contextualization*. Since the meaning of a node depends on the context where the node occurs, in the second phase we contextualize the interpretation of a node taking into consideration its ancestors. First, we perform sense filtering (i.e. we select the right WORDNET sense) and sense composition (i.e. we apply specific rules when the information deriving from the hierarchical structure is in conflict with the world knowledge provided in WORDNET). Then, combining the contextualized formulas of the ancestor nodes, we generate a logical form representing the meaning of the input nodes.

(iii) *Computation of the logical relation*. Since meanings are represented as logical forms, the problem of finding mappings between concepts can be transformed into a problem of logical deduction, i.e. a satisfiability problem where the possible mapping relations are tested and verified using a SAT solver.

3.1 Linguistic analysis of the labels

Concepts in classification hierarchies are described by labels, which in turn are composed of words and, possibly, separators between them. Labels are taken from a wide variety of linguistic expressions and can be single common words, such as *Music*, proper nouns, such as *Josef Strauss*, complex noun phrases, such as *Research Centers*, prepositional phrases, such as *Sociology of Religion*, verb phrases, etc. More complex labels can also contain conjunctions (e.g. *Ecological and Environmental Anthropology*), punctuation (e.g. *Clubs, Teams, and Societies*), and acronyms (e.g., *GIS*).

As a first step, we analyze the labels attached to the nodes from a linguistic point of view and generate a formula in description logic representing a first approximation of the meaning of the node. In this phase, nodes are interpreted as stand alone objects, which implies that the interpretation of a label will remain the same independently of the context and the position of the node to which it is attached. For example, the interpretation of a node labeled with *Submission* occurring under *ACL-02*, is equal to the interpretation of a node *Submission* occurring under *Senseval-2*.

Shallow parsing. The first step of the linguistic analysis consists of shallow parsing, i.e. dividing each label into syntactically correlated chunks. The chunker first tags each word with a part of speech; for example, in *Science Fiction and Horror* we have three nouns and a conjunction. Then, it identifies two noun groups (NGs), i.e. 'Science FICTION' and 'HORROR' (the syntac-

tic head is marked in small capitals), connected by a coordinating conjunction (1).

$$(1) [(Science)_n(FICTION)_n]NG \text{ (and)}_c [(HORROR)_n]NG$$

The output of the chunker is used to transform each label into a logical form. A noun group consisting of more than one word is interpreted as the conjunction of its elements. For instance, $[(Science)_n(FICTION)_n]$ is interpreted as $[Science \cap Fiction]$.

The relations between syntactic chunks are interpreted on the basis of the linguistic material connecting them:

- coordinating conjunctions and commas are interpreted as a disjunction;
- prepositions, like 'in' or 'of', are interpreted as a conjunction;
- expressions denoting exclusion, like 'except' or 'but not', are interpreted as a negation.

For example, *Science Fiction and Horror* is interpreted as a disjunction (2), which makes sense since under that node there might be both documents about 'science fiction' and documents about 'horror'; on the other hand, *Professional Photographers of America* and *Garments except Skirts* are examples of conjunction and negation respectively.

$$(2) [Science \cap Fiction] \cup [Horror]$$

Access to world knowledge. In order to perform the semantic interpretation of the labels we access the world knowledge provided in WORDNET. We use a multilingual version of WORDNET developed under the Meaning Project (Rigau et al. 2002).

When a word is found in WORDNET, all the senses of that word are selected and attached to the basic logical form. In the case of *Science Fiction and Horror*, for instance, WORDNET provides all three nouns, and so in the logical form we have the conjunction of the sets of senses of the lemmas (3).

$$(3) [science^* \cap fiction^*] \cup [horror^*]$$

We use the following notation: $science^*$ denotes the disjunction of all the senses of 'science' in WORDNET, while $science\#1$ and $science\#2$ indicate respectively sense number 1 and sense number 2 in WORDNET.

Multiword recognition. When two or more words in a label are contained in WORDNET as a single expression (i.e. a multiword), the algorithm selects the corresponding senses and substitutes the intersection between the senses of the single words with the senses of the multiword. For instance, WORDNET provides the expression 'science fiction', so the senses of the multiword are inserted in the logical form (4).

$$(4) [science_fiction^*] \cup [horror^*]$$

3.2 Contextual Interpretation

An interpretation of a concept in a CH as a stand alone object, however, is partial, as the meaning of a node depends on the context where the node occurs. Intuitively, the focus $f(c,H)$ of a node c belonging to the hierarchy H is the part of H that the user is required to visit in order to understand whether a document is in c . The contextualization of a node c gives a meaning to the node on the basis of the meaning of the nodes belonging to its focus (i.e. the ancestors of c with their direct descendants).

The logical form of a node is built combining the logical form of the node with the logical form of its ancestors through

intersection. As an example, let's take a concept hierarchy with the root *Soccer*, a descendant *Leagues*, and a further descendant *Clubs*. The logical form of the root is simply [soccer*], while the logical forms of *Leagues* and *Clubs* contain conjunctions, as shown in 5a and 5b respectively (the label attached to the node to which the logical form refers is highlighted in bold type).

- (5) a [soccer*] \cap [league*]
 b [soccer*] \cap [league*] \cap [club*]

Recognition of multi-level multiwords. The recognition of multiwords can also be performed on different contiguous levels. For instance, the multiword 'billiard player' is provided in WordNet, so it is inserted in the logical form of *Players*, which has *Billiards* as a parent node in the hierarchy in Figure 1.

- (6) a [sport*] \cap [billiards*] \cap [player*]
 b [sport*] \cap [billiard_player*]

Sense filtering. The context of a concept is taken into consideration to perform word sense disambiguation. We perform sense filtering exploiting both structural relations between labels and conceptual relations between words belonging to different labels.

Let L be a generic label and L^1 either an ancestor label or a descendant label of L and let s^* and s^{1*} be respectively the sets of WORDNET senses of a word in L and a word in L^1 . If one of the senses belonging to s^* is either a synonym, a hyponym, a hypernym, a meronym or a holonym² of one of the senses belonging to s^{1*} , these two senses are retained and all other senses are discarded.

As an example, imagine *Apple* (which denotes either a fruit or a tree) and *Food* as its ancestor; since there exists a hyponymy relation between apple#1 (denoting a fruit) and food#1, we retain apple#1 and discard apple#2.

Sense composition. As explained in Section 3, the presence of a coordinating conjunction makes the disjunction between noun groups within a label explicit, but we can also have **implicit disjunctions** between elements placed at different levels of the hierarchy (concepts with a disjoint descendant).

As an example, in the case of *Leagues*, with a parent node *Soccer* and a child *Clubs*, we have two conflicting interpretations: from the point of view of the hierarchical structure, *Clubs* denotes a subset of *Leagues* (being a child of it); on the other hand, from the point of view of the world knowledge provided in WORDNET, [club#2] (defined as 'a formal association of people with similar interests') and [league#1] ('an association of sports teams'), are disjoint because they have the same hypernym, i.e. association#1. In order to combine the two information sources, *Leagues* has to be reinterpreted as if it were *Leagues and Clubs* (7).

- (7) [soccer*] \cap [[league#1] \cup [club#2]]

When a concept is disjoint from one of its descendants, it has to be reinterpreted. More formally:

Let c and c' be two concepts, and let $c\#i$ and $c'\#j$ be two senses of c and c' respectively. We apply the following rule:

² A hyponym is a more specific concept, a hypernym is a more generic concept, a meronym is a part of a given whole, and a holonym is the whole of which a given concept is part.

³ $s\#k$ is disjoint from $t\#h$ if $s\#k$ belongs to the set of opposite meanings of $t\#h$ (if $s\#k$ and $t\#h$ are adjectives) or, in the case of nouns, if $s\#k$ and $t\#h$ are different hyponyms of the same synset.

- replace $c\#i$ with $c\#i \cup c'\#j$, if $c'\#j$ is disjoint from $c\#i$ and c is an ancestor of c' .³

Similarly, the negation can be marked by expressions like 'but not' or 'except', but we can have **implicit negations** in the case of elements belonging to different labels (inclusion relation between two siblings). For instance, in Google Web Directories we have *Sociology* and *Science* as sibling nodes classified under *Academic Study of Soccer*; from the point of view of world knowledge, sociology is a science (and in fact in WORDNET sociology#1 is a second level hyponym of science#2). As a consequence, the node labeled with *Science* has to be interpreted as if it were *Science except Sociology*.

When a concept is a hypernym or a holonym of a concept in another label on the same level, it has to be re-interpreted. More formally:

Let c and c' be two concepts, and let $c\#i$ and $c'\#j$ be two senses of c and c' respectively. We apply the following rule:

- replace the $c\#i$ with $c\#i \cap \neg c'\#j$, if $c\#i$ is either a hyponym or a meronym of $c'\#j$ and c and c' are siblings.

3.3 Computation of the logical relation

In the third step, we check whether a mapping relation, i.e. an equivalence, a more general or a less general relation, holds between the logical forms k and k' representing the meaning of the input nodes. To this aim, the task of finding a relation is transformed into a problem of propositional satisfiability (SAT), and then computed via a standard SAT solver. The SAT problem is built in two steps. First, the algorithm selects the portion T of the background theory relevant to the two logical forms, namely the semantic relations involving the WORDNET senses that appear in them. In the second phase, it computes the logical relations implied by T that represent semantic relations between the logical forms. The background theory T relevant for computing the relation between two formulas k and k' is obtained by transforming the WORDNET hierarchical relations between senses appearing in k and k' into a set of subsumptions in description logic according to the following rules:

- $c\#i \rightarrow c\#j$ (if $c\#i$ is a hyponym of $c\#j$)
- $c\#j \rightarrow c\#i$ (if $c\#i$ is a hypernym of $c\#j$)
- $c\#i \equiv c\#j$ (if $c\#i$ and $c\#j$ are synonyms)

The equivalence relation between k and k' (and consequently between the nodes whose meanings are represented by the logical forms) is checked by verifying that $k \subseteq k'$ and $k' \subseteq k$ are both implied by T . Similarly, the less [more] general relation between k and k' is checked by verifying that $k \subseteq k'$ [$k' \subseteq k$] is implied by T .

For example, the mapping between the source node *Broadcasting/Films/Science_Fiction_and_Horror* and the target node *TV/movies/science_fiction* is one of inclusion. The logical forms of the two nodes (8 and 9) and the logical relations implied by the background theory (10 and 11) are taken as input by SAT.

Through SAT we check for satisfiability the union of all the propositions (e.g. 10 and 11) and the negation of the implication between the logical forms 8 and 9. Since the check fails, a more general relation is computed between the two nodes; otherwise a similar procedure is followed for the other mapping relations.

- (8) [broadcasting#2] \cap [film#1]
 \cap [science_fiction#1 \cup horror#3]
 (9) [TV#1] \cap [movie#1] \cap [science_fiction#1]
 (10) film#1 \equiv movie#1
 (11) TV#1 \rightarrow broadcasting#2

4 Experiments and Discussion

The methodology described in Section 3 has been fully implemented in Java and is currently being intensively tested in a number of different application contexts. In this Section we present a large-scale evaluation of CTXMATCH performed on the Web Directories of Yahoo! (Yahoo!, 2003) and Google (Google, 2003), where documents are represented by Web page URLs.

Yahoo! and Google Web Directories have respectively fourteen and fifteen main categories (e.g. 'News & Media', 'Recreation & Sport', 'Health', 'Science', etc. in Yahoo!, 'Arts', 'Computer', 'Science', 'Society', etc. in Google) Each main category contains between a few thousand and tens of thousands of sub-categories and can be considered as the root of a CH.

A preliminary analysis has been performed on two pairs of sub-hierarchies, i.e. Google and Yahoo! 'Architecture' (under the main category 'Art') and 'Medicine' (under the main category 'Health'), whose sizes range between one hundred and one thousand nodes. Table 1 reports some linguistic data about the four CHs.

The labels attached to the nodes are generally short, with an average of 1.5-1.8 words per label. WORDNET's coverage is very high: between 88.7% and 95.5% of the words and lemmas occurring in the labels are found in WORDNET. Each lemma has on average between 3.2 and 4.6 senses, which makes the need for word sense disambiguation very important.

In order to evaluate automatically the logical relations computed by CTXMATCH we relied on the URLs classified under the two CHs, since it was not feasible to create a manual mapping between all possible pairs. The fact that Google and Yahoo! Web directories have been created manually guarantees, on the one hand, the high quality of the classifications and, on the other hand, makes them a good approximation of human judgement. The evaluation we present is based on the assumption that, given a source node and a target node belonging to different hierarchies, the higher the number of the documents (i.e. URLs) shared by the nodes, the higher the similarity between them.

	Architecture		Medicine	
	Yah.	Goog.	Yah.	Goog.
# Concepts	105	312	703	1,023
Average label repetition	1.0	1.3	2	1.8
# Words	170	521	1,231	1,549
# Words/label	1.6	1.7	1.8	1.5
WordNet's coverage (%)	95.5	91.5	88.7	91.4
Average polysemy	3.8	3.7	4.6	3.2
# Multiwords	11	45	51	116

Table 1: Analysis of the Architecture and Medicine sub-directories in Yahoo! and Google.

The evaluation was performed in four steps: (i) we identified the set D of documents classified in both CHs and selected the nodes containing at least one document belonging to this set; (ii) we established a correlation between the proportion of documents shared by source node and target node and the logical relation existing between them. The methodology for this was taken from Doan et al. (2002), who propose three formulas for calculating the similarity between nodes of CHs; (iii) we ran CTXMATCH on the selected nodes; and (iv) evaluated the mapping relations computed by CTXMATCH.

⁴ $P(X)$ denotes the probability of a randomly selected document from D being part of X .

Equivalence relation. The evaluation of the equivalence relation is based on the Jaccard coefficient (van Rijsbergen, 1979) calculated on two sets of documents: the set A of documents belonging to the common set of documents D classified under the source node, and the set B of documents belonging to D classified under the target node. According to (12)⁴, the similarity between the two sets is 1 if they contain the same documents and 0 if they are disjoint. Since in the Yahoo! and Google Web directories the number of documents shared by pairs of nodes is low and there can be different classifications of the same document due to human disagreement, we introduced an approximation factor e , so that an equivalence relation is judged as correct if the Jaccard coefficient ranges between 1 and $(1 - e)$, where e is empirically set to 0,1.

More [less] general relation. The *most-specific-parent* [*most-general-child*] measure (13) takes a value in the range $[0,1]$ when a node subsumes the other, so a more [less] general relation is correct if it ranges between 0 and 1.

$$(12) \quad \text{SIM}(A,B) = P(A \cap B) / P(A \cup B)$$

$$(13) \quad \text{MSP}(A,B) = \begin{cases} P(A|B) & \text{if } P(B|A) = 1 \\ 0 & \text{otherwise} \end{cases}$$

The results of the experiment are reported in Table 2, in terms of precision, recall, and F-measure obtained for the mapping relations returned by CTXMATCH. A baseline for the experiment was defined by considering a simple string match comparison among the labels placed on the path spanning from a concept to its root in the classification hierarchy (the results of the baseline are reported in brackets).

		Pr.	Re.	F
Arch.	equiv.	.75 (.80)	.08 (.05)	.14 (.09)
	more gen.	.94 (.96)	.38 (.37)	.54 (.53)
	less gen.	.84 (.90)	.79 (.51)	.81 (.65)
Med.	equiv.	.88 (.87)	.09 (.07)	.16 (.13)
	more gen.	.97 (.98)	.35 (.32)	.51 (.48)
	less gen.	.86 (.91)	.61 (.51)	.71 (.65)

Table 2: CTXMATCH results on Google and Yahoo! sub-directories 'Architecture' and 'Medicine'.

These results show that both the baseline and the CTXMATCH algorithm perform quite well. Not surprisingly, the baseline revealed itself as very precise, while CTXMATCH outperforms it with respect to recall. This confirms an important strength of CTXMATCH, namely that a content-based interpretation of contextual knowledge allows the discovery of non-trivial mappings.

As an example, in the sub-directory 'Medicine', the equivalence mapping between the concepts *Pharmacology/Psychopharmacology/Psychiatry* and *Psychiatry/Psychopharmacology* is found by CTXMATCH thanks to the recognition of a WORDNET hyponymy relation between *Pharmacology#1* and *Psychopharmacology#1*. An interesting mapping of inclusion is computed in 'Architecture' between *History/Periods_and_Styles /Gothic/Gargoyles* and *History/Medieval*; the source concept is less general than the target concept, which has been computed thanks to a WORDNET relation between *Medieval#2* and *Gothic#3*.

On the other hand, errors introduced during the linguistic analysis and contextualization of the labels may produce false positives. For instance, CTXMATCH erroneously suggests that *Medicine/Employment* (where 'employment' means 'job') is more

general than *Medicine/Optomety*; this is due to the wrong disambiguation of employment* as employment#4 defined in WORDNET as 'the act of using', which allows the algorithm to establish a relation with *Optometry*. Other errors arise when the algorithm is not able to perform sense filtering because of lack of relations between senses.

As far as sense composition is concerned, CTXMATCH works quite well with nouns denoting concrete objects, like buildings, but not as well with nouns denoting abstract concepts. Another limitation of the current system is that the use of the Alembic chunker does not permit the resolution of coordination ambiguities involving nominal compounds, and neither would the use of a more sophisticated parser, as this problem has received relatively little attention (Resnik, 1999), when compared to other aspects, like for instance prepositional phrase attachment.

5 Conclusions

We have addressed the problem of retrieving documents in classifications hierarchies as the problem of finding mappings between a source and a target hierarchy, in the perspective where a seeker with a given conceptualization of the world (i.e. the source CH) looks for certain documents classified according to a different conceptualization (the target CH).

The methodology we have proposed takes as input two nodes and returns a mapping relation between them. It does not consider the documents classified in the CH, thus allowing the retrieval of any kind of documents (e.g. text files, videos, images, etc.). The process of interpreting a label coincides with the progressive construction of a logical form in description logic, where predicates are WORDNET senses.

In the future we plan to work on a systematic analysis of the performance of the method with respect to the different steps and on the realization of a module for the discovery of the different kinds of relations between concepts, such as role, location, etc.

6 Acknowledgements

This work has been partially supported by the project EDA-MOK (Enabling Distributed and Autonomous Management of Knowledge), funded by the Provincia Autonoma di Trento with deliberation number 1060 on date 4/5/2001. We would like to thank the Edamok research team, and in particular the members of the "Context Matching Group" for useful discussion and invaluable feedback on this document.

References

- Agrawal, R. and Srikant, R.: On Integrating Catalogs. Proc. of the Tenth International World Wide Web Conference (WWW-2001), Hong Kong, China, May, 2001.
- Baader, F. and Nutt, W.: Description Logic Handbook. Pages 47-100, Cambridge University Press.
- Bouquet, P., Serafini, L. and Zanobini, S.: Semantic Coordination: A New Approach and an Application. Proc. of ISWC-03, Sanibel Island, Florida, USA, October, 2003.
- Calvanese, D., De Giacomo, G., Lenzerini, M.: A Framework for Ontology Integration. Proc. of SWWS International Semantic Web Working Symposium, Stanford University, USA, 2001.
- Daude, J., Padro, L., Rigau, G.: Mapping WordNets Using Structural Information. Proc. of ACL, Hong Kong, 2000.
- Day, D.S. and Vilain, M.B.: Phrase Parsing with Rule Sequence Processors: an Application to the Shared CoNLL Task. Proc. of CoNLL-2000 and LLL-2000. Lisbon, Portugal, September, 2000.
- Doan, A., Madhavan, J., Domingos, P. and Halevy, A.: Learning to Map between Ontologies on the Semantic Web. Proc. of WWW-2002, 11th International World Wide Web Conference, Honolulu, Hawaii, May, 2002.

- Fellbaum, C. (Ed.): WordNet: An Electronic Lexical Database. The MIT Press, Cambridge, US, 1998.
- Google. <http://directory.google.com/>, 2002.
- Ichise, R., Takeda, H., Honiden, S.: Integrating Multiple Internet Directories by Instance-based Learning. Proc. of IJCAI 2003, Acapulco, Mexico, August, 2003.
- Noy, N.F. and Musen, M.A.: Anchor-PROMPT: Using Non-Local Context for Semantic Matching. Proc. of the IJCAI-2001 Workshop on Ontologies and Information Sharing, Seattle, WA, August, 2001.
- Madhavan, J., Bernstein, P.A., Domingos, P., Halevy, A.Y.: Representing and Reasoning about Mappings between Domain Models. Proc. of AAAI 2002, Edmonton, Alberta, August, 2002.
- Resnik, P.: Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. Journal of Artificial Intelligence Research 11, 1999.
- Rigau, P., Magnini, B., Agirre, E., Vossen, P. and Carrol, J.: MEANING: a Roadmap to Knowledge Technologies. Proc. of the workshop "A Roadmap for Computational Linguistics", COLING-02, Taipei, Taiwan, 2002.
- Schulten, E., Akkermans, H., Botquin, G., Dörr, M., Guarino, N., Lopes, N. and Sadeh, N.: Call for Participants: The E-Commerce Product Classification Challenge. IEEE Intelligent Systems, 16-4, 2001.
- van Rijsbergen, C.J.: Information Retrieval. Butterworths, London, 1979. Second Edition.
- Yahoo! <http://www.yahoo.com/>, 2003.

Contact

Bernardo Magnini
ITC-irst, Via Sommarive, 18
38050 Povo (Trento), Italy
email: magnini@itc.it



Bernardo Magnini is Senior Researcher at ITC-IRST. He is involved in the Cognitive and Communication Technology division, where he coordinates a project on text processing technologies. His research interests are in the area of natural language processing and advanced information retrieval, with particular attention to question answering systems, word sense disambiguation and the application of NLP techniques to the Web scenario.



Luciano Serafini is a senior researcher of the Automated Reasoning Area of ITC-IRST. His main research interests are in the area of Distributed Knowledge Representation and Reasoning, with application to Knowledge Management, Information integration, Distributed Databases, and Semantic web. He is one of the main inventors of the Logic of Context, a logic for the representation of distributed knowledge.



Manuela Speranza graduated in foreign languages and literatures (major in linguistics) at the University of Trento in 2000. In the same year, she joined ITC-IRST to work on the development of a terminological linguistic resource. Since 2001 her research has focused on NLP techniques for distributed knowledge management.