KNOWLEDGE MEDIA INSTITUTE

Adaptive Named Entity Recognition for Social Network Analysis and Domain Ontology Maintenance

Tech Report kmi-04-30
December 2005

Jianhan Zhu, Alexandre L. Goncalves and Victoria Uren

The Open University

Adaptive Named Entity Recognition for Social Network Analysis and Domain Ontology Maintenance

Jianhan Zhu[1], Alexandre L. Goncalves[1,2] and Victoria Uren[1]

[1]Knowledge Media Institute, The Open University
Walton Hall, Milton Keynes
MK7 6AA, UK
{j.zhu, a.l.goncalves, v.s.uren, e.motta}@open.ac.uk

[2]Stela Group, Federal University of Santa Catarina
Florianópolis, Brazil
{a.l.goncalves}@stela.ufsc.br

November, 2004

Abstract

We present a system which unearths relationships between named entities from information in Web pages. We use an adaptive named entity recognition system, ESpotter, which recognizes entities of various types with high precision and recall from various domains on the Web, to generate entity data such as peoples' names. Given an entity, we apply a link analysis algorithm to the entity data for finding other entities which are closely related to it. We present our results to people whose names had been included for them to assess our findings. User feedback is analyzed by a statistical method. The results can be used to maintain a domain ontology. Our experiments on the Knowledge Media Institute (KMi) domain show that our system can accurately find entities such as organizations, people, projects, and research areas which are closely related to people working in KMi, and the results conform with the existing knowledge in our ontology and suggest new knowledge which can be used to update the ontology.

Categories and Subject Descriptors
I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*text analysis*; I.2.6 [**Artificial Intelligence**]: Learning—*knowledge acquisition*; H.3.3 [**Information System**]: Information Search and Retrieval—*search process*

Keywords: Named entity recognition, clustering, ontology

1. Motivation

As the Web penetrates into every corner of our lives, many activities in our real world have been recorded in a huge number of inter-linked Web documents. Web documents can serve as the mirror of what is actually going on, what has happened, and what things are related to each other, in our real world. Discovering these latent relations from the large number of Web documents can present a new way to organize information on the Web and help relieve the information overload problem.

Considering a corporate Web site, a domain which has thousands of or more Web documents, there are always new documents created, existing documents updated, and old documents removed. These documents often mention named entities such as people, organizations, and projects and describe what have happened between these entities. For example, in document 1, a person named “John Smith” met another person named “David Norman” from an organization “XYZ” company. In document 2, “John Smith” and “David Norman” joined “ABC” conference in a place. In document 3, “John Smith”, “David Norman”, and “Eddie Johnson” are projects members of “CDE” project. From documents 1, 2, and 3, we can find persons “David Norman” and “Eddie Johnson”, company “XYZ”, conference “ABC”, and project “CDE” which are related to “John Smith”.

However, when dealing with thousands of documents, we need to find a more efficient way to find latent relationships between these named entities. We can decompose the challenge into two sub problems. First, these named entities need to be accurately extracted from these documents. Second, given an entity, we need to use an algorithm to find other entities which are closely related to it based on a similarity measure.

In response to the first problem, there is extensive research on named entity recognition (NER) and information extraction techniques. However, due to the heterogeneous nature of the Web, an NER system needs to adapt to various domains on the Web in order to deal with problems such as homonyms, i.e., the same entity represents different things on different domains, e.g., “Magpie” is a bird on RSPB Web site but a project on KMi Web site, and variants, i.e., the same entity is represented in various forms, e.g., “Enrico Motta” can be represented as “E. Motta”, “Dr. Enrico Motta”, and “Prof. Enrico Motta”. In our previous work [Zhu et al. 2005], we presented an adaptive NER system called ESpotter, which adapts to various domains on the Web to recognize entities of various types with high precision and recall. A domain hierarchy is constructed from a link structure consisting of various domain on the Web to help disambiguate the same entity of different types of different domains and alignment different representations of the same entity using an entity similarity matrix. In the current study, we have used ESpotter, which is well suited for our aim, to process thousands of documents on a domain.

In response to the second problem, we propose to use clustering algorithms to find entities which are closely related to each other. Entities in a same document are deemed to co-occur with each other. We get a co-occurrence matrix whose rows are all the documents and columns are all the entities. Given this matrix, we can use different clustering algorithms to find entities which are closely related to each other based on criteria such as their co-occurrences and distributions, and compare performance of these algorithms.

We have used two ways to validate our finding. First, we present our findings to users for them to judge whether the most relevant entities to them are correct. Second, we can use a domain ontology, which is a representation of domain knowledge and often manually crafted and maintained, as the golden standard to corroborate our findings, i.e., locate entities in the domain ontology and find whether there is a direct or indirect relationship between related entities discovered by our method. At the same time, since domain ontology often fails to reflect ongoing activities due to its manual

maintenance shortcomings, our findings can help maintain the domain ontology by suggesting new entities, e.g., new project members, their relationships, e.g., a new PhD student for a PhD supervisor, and updating current entities, e.g., a PhD student start working as a research fellow, and their relationships, e.g., a research staff changed his line manager. We use the domain ontology to draw all possible relationships between two types of entities and present them to users in our evaluation, so the users can select the most appropriate relationship between himself/herself and other entities. Their responses are used to help maintain the domain ontology.

The rest of the paper is organized as follows. In Section 2, we give an overview of our system and its architecture. In Section 3, we discuss our work on using ESpotter to create entity data. In Section 4, we present our approach of using a link analysis algorithm to process the entity data for constructing a social network consisting of entities and their relationships with each other. The results are presented to users for evaluation. Our experimental results on KMi Web site are presented. In Section 5, we present our work on using the results for ontology maintenance. Our experimental results on KMi Web site are presented. In Section 6, we present related work. Finally, we conclude in Section 7.

2. Architecture

In the system architecture (Fig. 2.1), the domain of our study, **A**, is defined by a domain hierarchy obtained from clustering a link structure consisting of various domains on the Web as illustrated in our previous work [Zhu et al. 2004]. We find a set of Web documents, **B**, on the domain and use ESpotter, **C**, to process them for entity data, **D**. ESpotter can use the most relevant lexicon entries and patterns decided by the domain to perform NER with high precision and recall. ESpotter disambiguates the same entity of various types on different domains and aligns different representations of the same entity using a similarity matrix of the entities. The entity data, **D**, are presented in a document-to-entity matrix, whose rows are the documents and columns are entities found in these documents. The details are illustrated in Section 3.

We apply a link analysis algorithm, **E**, to the entity data in order to find most relevant entities measured by their similarity derived from the matrix. Entity relationships, **F**, are presented in a directed weighted graph consisting of entities as nodes and their relationships as directed links. Weights on the links show the strengths of the relationships. The details are illustrated in Section 4.

In order to add semantic meanings to these relationships, we extract a list of possible semantic relationships between two types of entities from a domain ontology, **H**. We evaluate our findings using two approaches. First, we ask users to verify a list of ranked entities which are relevant to himself/herself in **G**, and specify the type of semantic relationships between himself/herself and other entities. Second, we use the ontology to corroborate with our findings, i.e., see whether there are direct or indirect relationships between entities found in our method in the ontology. After user verification, domain knowledge which is not reflected in the current ontology is used to maintain the ontology, i.e., add new knowledge and modify out-dated knowledge. The details are illustrated in Section 5.

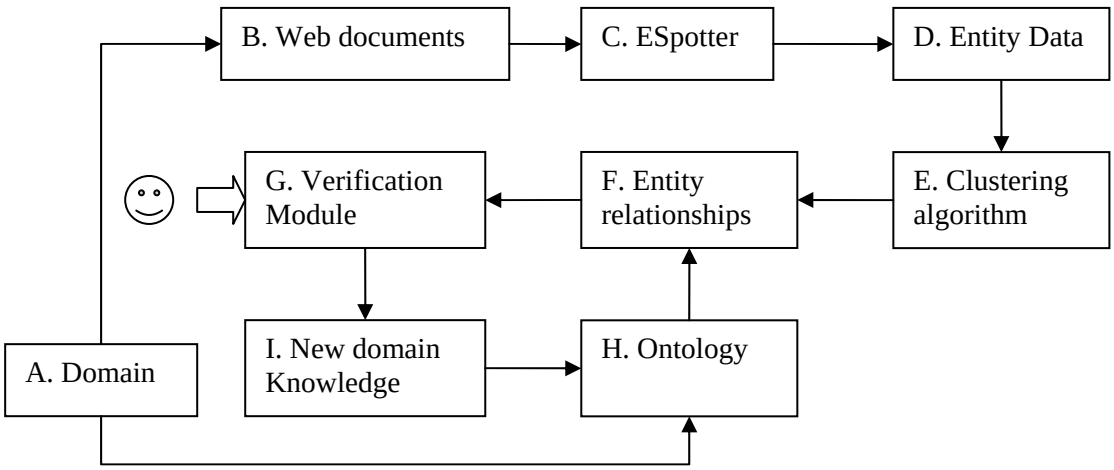

Fig. 2.1 System Architecture

3. ESpotter: Adaptive Named Entity Recognition

ESpotter adapts to various domains on the Web. ESpotter is a NER system capable of recognizing entities of various types from a large number of Web pages on various domains with high precision and recall efficiently. We propose a clustering algorithm to construct a domain hierarchy from a link structure consisting of domains. ESpotter uses the domain hierarchy for domain adaptation. ESpotter gets lexicon from domain knowledge such as a dictionary or ontology. ESpotter gets patterns manually crafted by human experts or on a domain using a wrapper learning method [Lixto]. We formalize lexicon entries and patterns as association rules. In estimating support and confidence of these association rules on domains of the domain hierarchy, we propose a search engine based query search method. The lexicon entry and pattern repository and domain adaptation module are modularized and allow user customization in NER for their special information search tasks on various domains. To our knowledge, ESpotter is the first system which provides a systematic way to solve domain diversity on a Web scale. ESpotter resolves ambiguity, i.e., the same entity of different types on different domains, and alignments, i.e., different representation of the same entity.

We define a domain hierarchy (such as in Figure 3.1) as a hierarchy which consists of domains on multiple levels and links between two domains on two adjacent levels. Domains are represented by their URIs. The root node is on the zero-th level. The higher the level of a domain is, the more general the domain is. A link from domain A to domain B means that A is the parent of B. If there is a link from domain A to domain B and a link from domain B to domain C, domain A is the ancestor of domain C.

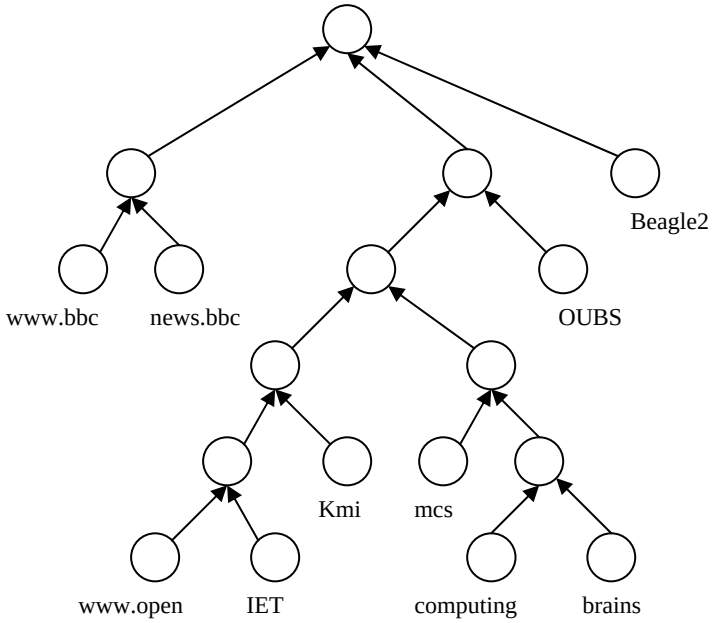

Figure 3.1 A Domain Hierarchy Defined on Domain URIs

Each lexicon entry or pattern is given support and confidence between 0% and 100% on each node in the hierarchy. The higher both the values, the stronger the lexicon entry or pattern.

As the output of ESpotter, we get a document-to-entity matrix, whose rows are documents and columns are entities such as in Table . D1 to D5 are five Web documents. E1 to E7 are seven entities of different types. T1 to T3 are three different entity types. The number in each entry of the matrix is the number of occurrences of the corresponding entity in the corresponding document, and zero for the empty entry.

Table 3.1 An Example of a Document-to-Entity Matrix constructed by ESpotter

	(E1,T1)	(E2,T1)	(E3,T1)	(E4,T2)	(E5,T2)	(E6,T2)	(E7,T3)
D1	2	1		1		1	1
D2	2		1	1	1	1	1
D3		1	1	1		1	1
D4	3	2	1		1		1
D5	1	1	1	1	1		1

4. Methodology: Toward Relationship Identification for Social Network Analysis

4.1 Link Analysis

Aiming to achieve the objectives, we developed a methodology in order to extract useful information. The steps used are described in more details below. Fig. 2.1 presents the system architecture as well as all the requited phases of the methodology.

The first step consists in the composition of an adjacency matrix (Entity-to-Entity) based on our database regarding just the concepts used in this study (Organization, Person, Project and Research Area). The number of entities extracted during the matrix composition was 2164, being 726 Organization, 1002 Person, 21 Project, and 415 Research Area. During this process a co-occurrence matrix is produced taken into account the frequency to each entity pairs <*E*1,*E*2>, the average distance intra-document, and the probability of two concepts together over the whole corpus. For example, analyzing Table 3.1, E1 would have a relationship with all the other entities and concepts.

We have tried some equations aiming to normalize the relationships such as *tfidf* and the *tflog*. The equation *tfidf* showed poor results yielding a wrong classifications regarding that important concepts such as KMi and Open University did not appear in the first positions. Although this equation is quite important in a wide range of applications it seems to be not suitable to measure the relation strength among concepts embedded in documents. The *tflog*=1+log(*tf*), where *tf* means the frequency, just reduce the importance by considering, for example, that a term/concept with frequency 3 is not 3 times more relevant than one occurring once. Two other aspects are important, the distance between the entities intra-document and the probability of them appear together over the whole corpus.

The average distance takes into account the offsets of both entities calculating the minimum offset between each possible relation intra-document normalized using the following classes: (0,5]=1, (5,10]=2, (10,20]=3, (20,40]=4, (40,75]=5, (75,125]=6, (125,200]=7, (200,300]=8, (300,500]=9, otherwise 10. The equation is defined as:

$$\bar{d}(E1,E2) = \sum range(\max_{offset}(E1,E2) - \min_{offset}(E1,E2) - len(E1,E2)), \quad (4.1)$$

where, function *range*() is based on above description, *len*(*E1*,*E2*) means the entity string length with maximum offset. For example, regarding entities having offsets as follows *E1*={350,2500,4500} and string length of 15, and *E2*={400,3200} and string length of 12, the result is:

$\bar{d}(E1,E2)$ = (range(400-350-12)+ range(3200-2500-12)+ range(4500-3200-15))/3 = 8

We have take into account that strong relationships must be included across the whole corpus, meaning that the more intersections are found, the more relevant is the relationship between them. In this work we are using the Resnik's [15] method for noun probability to compute the Entity probability as relative frequency:

$$\hat{p}(E1,E2) = \frac{freq(E1,E2)}{N}, \quad (4.2)$$

where, *freq*(*E*1,*E*2) is the frequency summation of *E*1 and *E*2 when occurring together over the total number of documents *N*. Thus, the weight of a pair of entities is

calculated through the summation of all the intra-document relationships normalized through the probability extra-document, as defined:

$$w(E1,E2) = \sum \left(\frac{tf \log_{E1} * tf \log_{E2}}{\bar{d}(E1,E2)} \right) * \hat{p}(E1,E1) \quad (4.3)$$

where, *ftlog* is used to normalizes the entity frequencies, $\bar{d}(E1,E2)$ means the average distance and $\hat{p}(E1,E2)$ the entity probability. Table 4.1 presents an example of adjacency matrix with the calculated weights. .

Table 4.1. The adjacency matrix (Entity-to-Entity) calculated from the Document-to-Entity matrix in Table 3.1

Entity	E1	E2	E3	E4	E5	E6	E7
E1	-	2.422	1.196	0.285	4.955	1.620	0.897
E2	2.053	-	21.491	0.135	1.321	1.672	0.125
E3	0.809	23.050	-	0.065	0.733	0.598	0.022
E4				-			
E5					-		
E6						-	
E7							-

The final matrix provides a direct relationship scenario among entities. It enables to have insights about the entity organization in a specific organization, for example. However in order to establish latent relationships it is needed to cluster these entities.

4.2 Latent Relationships

Using the Entity-to-Entity matrix is composed a set of vectors considering just lines in which the entity of concept type Person has occurred forming a more complex vector space. Table 4.2 presents relationships extracted from E1 entity.

Table 4.2. Example of vector extract from E1 considering the three main concepts

Concept	Related Entity	$W(E1,E_n)$
Organization	E2	2.422
Organization	E3	1.196
Organization	E4	0.285
Person	E5	4.955
Person	E6	1.620
Person	E7	0.897
Project	E8	0.123
Project	E9	0.051
Project	E10	0.048
Research Area	E11	0.373
Research Area	E12	0.362
Research Area	E13	0.290

4.3 User Evaluation

The last phase is important to the whole process and aims to collect the user feedback about the automatic ranking and relations. The evaluations were carried out by 15 people that evaluated the suggested ranking regarding the concepts used in our study through an application. An example of the user evaluation interface (in Fig. 4.1) presents the application scenario based on Person concept. Similar view is gotten by using Organization and Project concepts. Using it is possible to change the suggested ranking by selecting the Order object, as well as it is possible to inform some kind of relation with the related entity. In order to better evaluate and understand the proposed ranking, it was asked to users to analyze the news associated with each entity. It is done by selecting the link over each entity. Comments about ranking, relation and possible misclassification can be made using the appropriate field available on the interface.

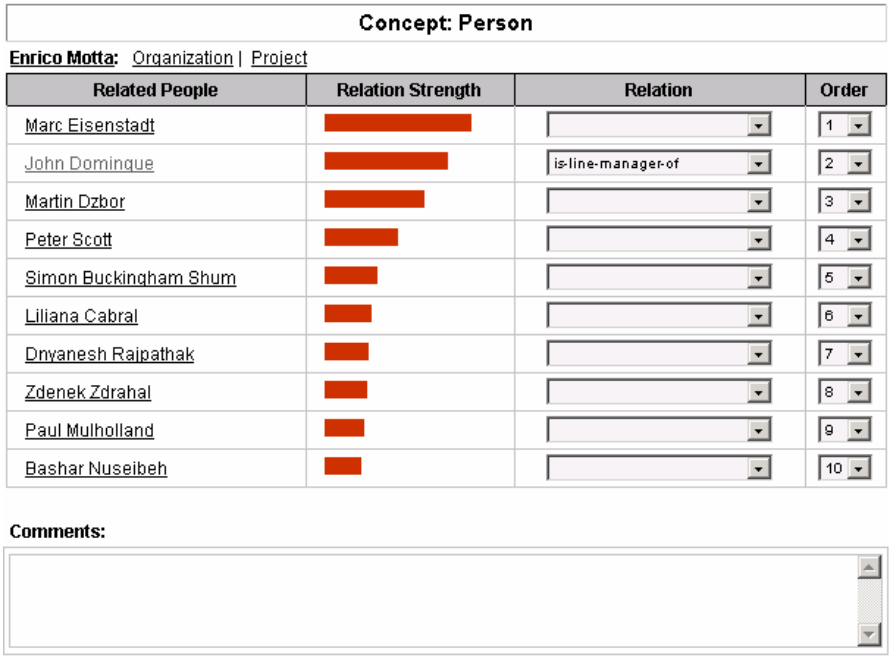

Fig. 4.1 Application to collect the user feedback

5. Qualitative Comparison of Social Network and KMi Basic Portal Ontology

It is envisaged that one use of the data returned by the social network system will be to feed an ontology maintenance system. Therefore a comparison was made between the listings produced by the system and the KMi basic portal knowledge base. This was done manually in the first instance using OCML to explore the ways in which comparisons could be made. For the initial analysis, three people were chosen, representing different roles within the KMi community. These were a professor (Marc Eisenstadt), a research fellow (Martin Dzbor) and an external PhD student (Al Selvin). We looked at relations between people and projects, and people and people. Since the only organizations represented in the knowledge base were KNOWLEDGE-MEDIA-INSTITUTE-AT-THE-OPEN-UNIVERSITY and CNM the people – organization links were not queried. We note that this is a fertile area for populating the knowledge base.

For querying the links between people and organizations a new relation WORKS-ON-PROJECT was created which bundles together all the possible relations between people and projects.

```
(def-relation WORKS-ON-PROJECT (?person ?project)
        :iff-def (or (has-project-member ?project ?person)
                (has-contact-person ?project ?person)
                (has-project-leader ?project ?person)))
```

We also created a "share interest" relation, which was suggested by one of the subjects as an extra option for relations. The later could be modelled by overlap of the has-research-interest slot for people and the addresses-generic-area-of-interest slot for projects which use the same fillers.

```
(def-relation SHARE_INTEREST (?person ?project ?interest)
        :iff-def (and (has-research-interest ?person ?interest)
        (addresses-generic-area-of-interest  ?project ?interest)))
```

For querying the links between people and people three new relations were created: COLLABORATES-WITH, WORKS-ON-PROJECT and PEOPLE-SHARE-INTEREST. COLLABORATES-WITH uses WORKS-ON-PROJECT to identify people who work on the same project.

```
(def-relation COLLABORATES-WITH (?person1 ?person2)
        :iff-def (and (works-on-project ?person1 ?project)
                (works-on-project ?person2 ?project)))
```

COLLABORATES-WITH was not used directly for queries but was incorporated in the main relation WORKS-WITH which bundles together all the possible relations for people.

```
(def-relation WORKS-WITH (?person1 ?person2)
        :iff-def (or (has-supervisor ?person1 ?person2)
                        (has-supervisor ?person2 ?person1)
                (has-line-manager? person1 ?person2)
                        (has-line-manager? person2 ?person1)
                (COLLABORATES-WITH ?person1 ?person2)))
```

PEOPLE-SHARE-INTEREST is very similar to the SHARE-INTEREST relation for people and projects.

```
(def-relation PEOPLE-SHARE-INTEREST (?person1 ?person2 ?interest)
        :iff-def (and (has-research-interest ?person1 ?interest)
                (has-research-interest ?person2 ?interest) ))
```

Experimental Evaluation

6.1 Semantic Network Analysis

We evaluated our methodology and proposed model by comparing the ranking provided automatically with users' feedback. Spearman's rank correlation coefficient [Powell and French, 2003] was used in this analysis, formally defined as:

$$R = 1 - \frac{6\sum d^2}{n^3 - n} \tag{6.1}$$

where $\sum d^2$ is the sum of the squares of the rank differences, and n the number of concepts ranked, being in our case, 10 the maximum value. Regarding $-1 \leq R \leq 1$, $R = 1$ indicates two rankings in perfect agreement and $R = -1$ in perfect disagreement. We used our PlanetNews dataset[1] to evaluate our approach and present the results to 15 users for evaluation. Table6.1 presents the user evaluation feedback as the correlation to Organization, Person and Project.

Table 6.1. Spearman's rank correlation between automatic ranking and user trial. The null values in project column mean that there are not projects associated with the person

Users	Organization (*R*)	Person (*R*)	Project (*R*)
Person 1	-1	0.4788	-0.4
Person 2	-0.1636	0.8909	-0.4
Person 3	-0.1273	0.8545	-0.2571
Person 4	-0.0286	1	-
Person 5	0.2381	0.9152	0.5
Person 6	0.3095	0.4762	-
Person 7	0.3571	-0.0545	1
Person 8	0.4909	0.7939	0.4182
Person 9	0.6121	0.8929	1
Person 10	0.9394	0.4	0.8
Person 11	0.9879	1	0.3500
Person 12	0.9879	1	0.9643
Person 13	0.9879	0.9879	1
Person 14	1	0.8667	1
Person 15	1	0.9879	1

In order to measure the ranking correlation significance we applied the *t* test when there were at least 10 related concepts, using Equation 5.2 with *n* -2 degrees of freedom.

$$t = R\sqrt{\frac{n-2}{1-R^2}} \tag{6.2}$$

Through this test was adopted an *R* cutoff point aiming to determine the degree of agreement between the suggested ranking and the evaluated ranking. The probability level desired was 5% of significance achieving *R* = 0.65. Values below this point are going to be classified as "Do not agree" and values above it as "Agree". Table6.2 presents the summarized rank correlation according the defined classes.

Table 6.2. Classes based on *t* test to summarize the rank correlation

[1]We used 146 news stories published after 2001 for this analysis.

Class	Organization	Person	Project
Agree	40%	73%	54%
Do not Agree	60%	27%	46%

Analyzing the table is possible to verify considerable levels of significance in both Person and Project concepts, with 73% and 54% respectively. The result presented at Person concept shows that people working together tend to have their activity registered in somehow. In our database it happens mainly through publications and technical reports. Despite the achievement of 54% in Project concept, this figure reflects the lack of information. Projects are in general accomplished by a set of people and take some time to be finish. In this sense, people use to work in least projects than they use to keep professional relationship. On the other hand, looking back to the database is noticed that just 2 users, changed the project classified as the first to positions further second. In order to get a representative database considering this concept a longer period must be considered.

Regarding Organization concept is noticed that just 40% agree with the automatic ranking. It can be justified by the lack of the main organization where these people are developing their work. Despite it and taking into account just the top five associated concepts there are few modifications in the ranking and generally the three most relevant still remain among the top five.

6.2 Qualitative Comparison of Social Network and KMi Basic Portal Ontology

From relations from the ontology a binary analysis was performed to determine whether any relation exists in the ontology for a relation identified by the social network system.

6.2.1 Results for PEOPLE/PROJECTS

These relations were identified using the WORKS-ON-PROJECT relation and the SHARE-INTEREST relation for people and projects.

For Marc Eisenstadt we determined that, with the exception of BuddySpace, there is no WORKS-ON-PROJECT link in the knowledge base between him and the projects he was associated with based on the Planet News stories. However he does have a SHARES-INTEREST link with 7 out of the 10 projects. This may be because there are 16 interests specified for Marc in the ontology, causing a high probability of a hit. Marc specified relationships with three projects: Buddy Space, CoAKting, and ClimatePrediction. Marc did not attempt to reorder his relations but his three edits suggest that ClimatePrediction at CoAKTinG should be further up the ranking to correctly reflect the situation.

Table 6.3 Results for Marc Eisenstadt

Project	WORKS-ON-PROJECT	SHARE-INTEREST	Project	Response
BuddySpace	1	1	1	project leader
Magpie	0	1	1	other
Compendium	0	1	1	other
AKT	0	0	1	other
Rostra	0	1	1	other
CIPHER	0	1	1	other
CoAKTinG	0	0	0	project member
D3E	0	1	1	other
IRS	0	0	1	other
ClimatePrediction	0	1	1	project member
TOTAL	1	7	9	10 changes

For Martin Dzbor we found WORKS-ON-PROJECT relations in the knowledge base for all the projects for which he specified a relation. We also found 7 SHARE-INTEREST relations, despite the fact that Martin had only specified 8 interests. It is possible that the interest slots are too densely populated to be informative about relations. Martin reranked his entries and reported that, while the first 2 were good the rest were fairly poor. It can be seen that he moved projects with which he turned out to have a relation specified in the ontology to the top of the ranking.

Table 6.4 Results for Martin Dzbor

Project	WORKS-ON-PROJECT	SHARE-INTEREST	Project	Response	Rerank
Magpie	1	1	1	Contact person	1
BuddySpace	1	1	1	Project member	2
Compendium	0	1	1		8
Rostra	0	1	1	other	7
IRS	0	0	1	other	5
AKT	1	0	1	Project member	4
D3E	0	1	1		9
CoAKTinG	0	0	0	other	6
Scholonto	0	1	1		10
ClimatePred.	1	1	1	Contact person	3
TOTAL	4	7	9	7 changes	

For Al Selvin we determined that there is no entry for Al himself. Consequently none of the relations between him and the projects his name occurs with in Planet News exist. One of the projects (CoAKTinG) also turned out to be missing. Al did not rerank his entries but his responses suggest that the social network has done a reasonable ranking here, putting the project Al leads (Compendium) above the one for which he is a member (CoAKTing).

Table 6.5 Results for Al Selvin

Person	Project	Person	Project	Response
Al Selvin	Compendium	0	1	Project leader
Al Selvin	CoAKTinG	0	0	Project member
Al Selvin	BuddySpace	0	1	Other

6.2.3 Results for PEOPLE/PEOPLE

These relations were identified using the WORKS-WITH relation and the PEOPLE-SHARE-INTEREST relation.

For Marc Eisenstadt we found that one of the people he was associated with did not exist (Bashar Nuseibeh). He shared a WORKS-WITH relation with five of the remaining people but specified a relation for all nine. However many of these were the vague relation has-similar-research-interest. Only two of the WORKS-WITH relations were specified more closely than this. Interestingly two of the people with whom Marc believes he shares interests were not picked up by the SHARE-INTEREST relation (Paul Mulholland and Zdenek Zdrahal). On inspection it transpired that neither has any research interests specified in the ontology, which is probably an omission. Once again Marc did not reorder his listing so we cannot judge how good the ranking was from this.

Table 6.6 Results for Marc Eisenstadt

Person	WORKS-WITH	PEOPLE-SHARE-INTEREST	Person	Response
Enrico Motta	0	1	1	has-line-manager
Martin Dzbor	1	1	1	has-similar-research-interest
Simon Buckingham Shum	1	1	1	has-similar-research-interest
Jiri Komzak	1	1	1	is-line-manager-of
Paul Mulholland	0	0	1	has-similar-research-interest
Peter Scott	1	1	1	has-similar-research-interest
John Domingue	0	1	1	has-similar-research-interest
Zdenek Zdrahal	0	0	1	has-similar-research-interest
Yanna Vogiazou	1	1	1	is-supervisor-of
Bashar Nuseibeh	0	0	0	other
TOTAL	5	7	9	

For Martin, we found that he WORKS-WITH 7 of the people on his list which is a promising result, particularly as two of the remaining people do not exist in the knowledge base. However the third person, Zdenek Zdrahal, was given a specific relation is-supervisor-of by Martin. This alerted us to the fact that there is not a single instance of this relation in the knowledge base, suggesting a systematic omission. We also found that, with 8 research interests of his own, Martin shares an interest with 7 of the people on his list, supporting the view that this is not a very discriminating relation. Because there were so many hits it was difficult to draw any conclusions about how good the ranking was based on the binary analysis, but Martin commented that “first two-three people quite good, the rest not very convincing” so it seems there is room for improvement.

Table 6.7 Results for Martin Dzbor

Person	WORKS-WITH	PEOPLE-SHARE-INTEREST	Person	Response	Rerank
Marc Eisenstadt	1	1	1	Has-similar-research-interest	1
Enrico Motta	1	1	1	Is-line-manager-of	2
Jiri Komzak	1	1	1	Has-similar-research-interest	3
Simon Buckingham Shum	1	1	1	other	8
John Domingue	1	1	1	Has-similar-research-interest	4
Liliana Cabral	1	1	1	Has-similar-research-interest	6
Dnyanesh Rajpathak	1	1	1	Has-similar-research-interest	7
Bob Spicer	0	0	0	other	9
Zdenek Zdrahal	0	0	1	Is-supervisor-of	5
Al Selvin	0	0	0	other	10
TOTAL	7	7	8	10	

6.2.4 Outcomes

The results suggest that the social network can expose omissions of instances from the knowledge base, such as the absence of most of the organizations, Al Selvin, Bashar Nuseibeh, the CoAKTing project etc.. It can also show up deficiencies in the slot filing such as the lack of any research interests for Paul Mulholland and Zdenek Zdrahal.

The binary ontological analysis method used so far needs to be elaborated. In a close-knit community like KMi people share so many interests and collaborate together so is so many ways that giving binary results from the WORKS-WITH and PEOPLE-SHARE-INTEREST relations is uninformative. A measure which gave, for example, the proportion of the main person's interests which they share with the query person or the number of projects on which they collaborate might tell us more about the relation between them.

The ranking should be improved, if possible, to supply people with closer matches. This may happen as a by product of the proposed repeat of the experiment with a larger sample of data.

6. Related Work

Named entity recognition is a well studied area [Cunningham 2000; Grover et al.]. We used ESpotter, which improves on traditional NER systems by adapting lexicon entries and patterns to various domains on the Web for high precision and recall on documents on these domains.

Kruschwitz [2003] proposed to construct a network consisting of terms based on their occurrences in a collection of documents on a certain domain. The network is used to improve the quality of user query. WordNet [2000] consists of typical English terms and their relationships with each other. Heylighen [2001] proposed applications of associative networks, which learn associations through Hebbian-style rules either by

measuring co-occurrence of words in text or patterns of usage, to problems of ambiguity and meaning in language.

Many methods have been proposed to extract domain terminology or word associations from texts and use this information to build or enrich an ontology [Maedche 2002; Morin 1999; Vossen 2001] Missikoff et al. [2002] proposed the OntoLearn system which supports the construction and assessment of a domain ontology for intelligent information integration within a virtual user community.

Spearman's Rank Correlation [Borkowf 2000] is a method used for calculating correlation between variables, when the data does not follow the normal distribution. Link analysis is closely related to clustering. Visvimo [2000] cluster search results returned by typical search engines such as Google. Dumais and Chen [2000] proposed to cluster search results by Web page contents.

7. Future Work

8.1 Semantic Network Analysis

We are now using other clustering algorithms such as SVD and LSI to construct the semantic network and compare their performance with the link analysis algorithm used. The advantage of SVD and LSI is that they can find strong relationships between entities which are not directly linked in our social network. The PlanetNews data is biased and we are using the whole KMi Website consisting of 1863 documents for our study. We plan to integrate shallow language understanding techniques to help associate entities in the same document with each other and assign semantic relationships between entities.

8.2 Requirements for an ontology maintenance application

From this manual analysis the following requirements for an ontology maintenance system can be determined

Approximate match of instances - The problem of matching, e.g., "IRS" found by ESpotter against Internet-Reasoning-Service in the knowledge base was solved in this manual analysis by human domain knowledge. For an automatic or semi-automatic version some kind of approximate matching is required.

Derived relations – in this analysis the possible relations were constructed on a domain specific basis. A general purpose tool would need to be able to derive possible connections automatically from the ontology.

Complexity of feedback – We used a binary approach for this study, noting only whether any relation of the given type was present. As the results with the PEOPLE-SHARE-INTEREST connection shows this can be uninformative. Therefore more complex feedback based on counting the total number of a particular kind of connection between two instances may be required.

We note that the second of these requirements has already been partially solved by the Aqualog system which can derive single link relations (but not 2 link ones like SHARES-INTEREST). The first requirement for approximate matching is currently being addressed in Aqualog with the improvement of its string-matching capability.

The verification module of the Semantic website is also incorporating improved string matching of instances. It therefore seems sensible to try to combine these activities.

8. Conclusion

In this paper, we propose a novel approach of extracting named entities from a number of documents on a domain for constructing a semantic network consisting of these named entities and their relationships. We validate the semantic network by users in the domain and the domain ontology. New verified knowledge is used to maintain the domain ontology. Our experiments on KMi domain show that relationships between people working in KMi with other entities in the semantic network constructed by our method match well the opinions of these people, and the new entities and relationships are helpful in our ontology maintenance.

9. Acknowledgements

This research was partially supported by the Designing Adaptive Information Extraction from Text for Knowledge Management (Dot.Kom) project, Framework V, under grant IST-2001-34038 and the Advanced Knowledge Technologies (AKT) project. AKT is an Interdisciplinary Research Collaboration (IRC), which is sponsored by the UK Engineering and Physical Sciences Research Council under grant number GR/N15764/01. The AKT IRC comprises the Universities of Aberdeen, Edinburgh, Sheffield, Southampton and the Open University. Alexandre L. Goncalves is supported by CNPq, Brazil, with a doctoral scholarship.

Reference:

1. Chaomei Chen: Structuring and Visualising the WWW by Generalised Similarity Analysis. Hypertext 1997: 177-186
2. Gary William Flake, Steve Lawrence, C. Lee Giles, Frans Coetzee: Self-Organization and Identification of Web Communities. IEEE Computer 35(3): 66-71 (2002)
3. Jianhan Zhu, Victoria Uren, Enrico Motta (2004) *ESpotter: Adaptive Named Entity Recognition for Web Browsing and Search*. Draft
4. Jianhan Zhu, Victoria Uren, and Enrico Motta. *ESpotter: Adaptive Named Entity Recognition for Web Browsing*. To appear in Proc. of Workshop on IT Tools for Knowledge Management Systems at WM2005 Conference, Lecture Notes in Computer Science, Springer, Kaiserslautern, Germany, April 11-13, 2005.
5. K. Church and P. Hanks, "Word Association Norms, Mutual Information, and Lexicography", *Computational Linguistics*, vol. 16, issue 1, 1990, pp. 22-29.
6. S. Ross, "A First Course in Probability", Macmillan, 1976.
7. Craig B. Borkowf. *Computing the Nonnull Asymptotic Variance and the Asymptotic Relative Efficiency of Spearman's Rank Correlation*. Computational Statistics & Data Analysis, vol. 39, issue 3, 2002, pp. 271-286.
8. Allison L. Powell and James C. French. *Comparing the performance of collection selection algorithms*. In: ACM Transactions on Information Systems (TOIS), vol. 21, issue 4, 2003, pp.412-456.
9. Kruschwitz, U. An Adaptable Search System for Collections of Partially Structured Documents. IEEE Intelligent Systems, 18(4): 44-52, July/August 2003.
10. Michele Missikoff, Roberto Navigli, Paola Velardi: Integrated Approach to Web Ontology Learning and Engineering. IEEE Computer 35(11): 60-63 (2002)

11. A. Maedche, “Emergent Semantics for Ontologies--Support by an Explicit Lexical Layer and Ontology Learning,” IEEE Intelligent Systems, 2002, http://wim.fzi.de/wim/publications/entries/1010141835.
12. E. Morin, “Automatic Acquisition of Semantic Relations between Terms from Technical Corpora,” Proc. 5th Int’l Congress Terminology and Knowledge Extraction (TKE-99), TermNet, Vienna, 1999, pp. 268-278.
13. P. Vossen, “Extending, Trimming and Fusing WordNet for Technical Documents,” NAACL 2001 Workshop WordNet and Other Lexical Resources, 2001; http://citeseer.nj.nec.com/447335.html.
14. Heylighen F. (2001) Mining Associative Meanings from the Web: from word disambiguation to the global brain. In Pro. of the International Colloquium: Trends in Special Language & Language Technology, R. Temmerman & M. Lutjeharms (eds.) (Standaard Editions, Antwerpen), p. 15-44.
15. Philip Resnik: 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: 95-130 (1999).