Mining Web Data for Competency Management

J.Zhu[1], A.L.Gonçalves[2], V.S.Uren[1], E.Motta[1], R.Pacheco[2]
[1]*Knowledge Media Institute, The Open University, Milton Keynes, United Kingdom*
{e.motta, v.s.uren, j.zhu}@open.ac.uk
[2]*Stela Institute, Florianópolis, Brazil*
{a.l.goncalves, pacheco}@stela.ufsc.br

Abstract

We present CORDER (COmmunity Relation Discovery by named Entity Recognition) an un-supervised machine learning algorithm that exploits named entity recognition and co-occurrence data to associate individuals in an organization with their expertise and associates. We discuss the problems associated with evaluating unsupervised learners and report our initial evaluation experiments.

1. Introduction

At a time in which organizations increasingly regard the knowledge and skills of their employees as their most valuable resource, competency management, knowing who knows what, has become a critical activity. Equally important is knowing who knows whom, both outside and inside the organization, so that project teams with the right mix of skills, contacts and experience of working together can be assembled.

We argue that documents are a primary resource for discovering information about people's skills and associations. Text based approaches have already been used in some specialist domains, for example to create a database about the competencies of expert witnesses [1]. We propose to use text documents in a more general scenario and to concentrate on finding the relations between entities of several kinds rather than classifying experts against a taxonomy of skills. The documents used may be intended to summarize competency information, such as the large collection of Brazilian researchers' curricula vitae held on the Lattes Platform (http://lattes.cnpq.br/historico.jsp), but they might equally be ordinary documents, such as Web pages and reports which reflect day to day activity within the organization.

We propose tackling competency discovery from documentary resources using an unsupervised web content mining algorithm which we call CORDER (COmmunity Relation Discovery by named Entity Recognition). Named Entity Recognition (NER) is used as a preliminary step to identify named entities (NEs) of interest, such as people's names, organization names and knowledge areas, thus partially tackling the problem of un-structured Web data identified by some of the earliest writers on web-mining [2]. The output of the CORDER algorithm is a matrix of entities. In a competency management scenario the target dimensions of this matrix is people, for example employees or researchers. The other dimension is the parameters against which competency is being assessed, which might be subject domains, contacts, organizations with which a person has collaborated or projects they have worked on. The values in the matrix are the relation strength calculated by CORDER of the relation between a person and a given parameter.

CORDER builds on work such as DIPRE [3], Snowball [4] and KNOWITALL [5], but, because it uses co-occurrence rather than relatively rare patterns for discovering relations, it can discover relations in collections smaller than the whole Web, making it suitable for corporate intranets. CORDER has similarities to the relation discovery method of Hasegawa et al. [6] which clusters pairs of NEs according to the similarity of context words between them. Their method works well on newspaper text, which usually consists of well-formed sentences. The advantage of the co-occurrence method we use is that it is general enough to detect relations in inhomogeneous text where relations may not be explicitly specified by context words. Their method also does not address ranking relations in terms of relevance.

This paper is organized as follows. In Section 2 we describe the CORDER algorithm. In section 3 we discuss issues concerning the evaluation of unsupervised machine learning algorithms in general. Sections 4 and 5 present two evaluation studies, one based on expert evaluation of CORDER's results and the other comparing the results to a quantitative benchmark. Finally in Section 6 we conclude and describe some of our on-going work applying and refining the algorithm.

2. CORDER

CORDER discovers relations by identifying co-occurrences of NEs in text. This approach is based on the intuition that if an individual has expertise in an area his/her name will be associated with key terms about that area in many documents. Similarly if two individuals often work together we expect to see their names associated. In general, we assume that NEs that are closely related to each other tend to appear together more often and closer.

The process CORDER follows comprises the steps of:

1. *data selection*, in which the Web pages that will represent the organization are identified,
2. *named entity recognition*, in which the pages are preprocessed, and
3. *relation strength and ranking*, in which co-occurrence data is processed and the relation strengths of associated NEs with the target are established.

We describe these steps below, concentrating on relation strength.

2.1 Data Selection

We find Web pages from an organization's Web site using a Web spider. Web pages, which contain noisy data, e.g., out-dated information and irrelevant information, may be removed. Web pages which are linked from the Web site may be taken into account if they contain relevant information.

2.2 Named Entity Recognition

A named entity recognizer is used to recognize people, projects, organizations and research areas from the Web pages. We use ESpotter [7], an NER system which employs standard NER techniques, because it provides methods for rapidly adapting its lexicon and patterns to different domains on the Web. Automated Google searches are used to estimate the number of times a pattern or lexicon occurs on the Web in general and on pages with a URL associated with the domain. These are used to estimate the probability of particular patterns on the domain of interest. Recall and precision can then be controlled by adjusting a threshold parameter to select which patterns should be used on a given domain. Fine adjustments can be made by the user for individual patterns. This combination of automatic probability estimation and manual refinement allows ESpotter NER to be optimized for a particular organization's pages without a long training process.

Variants of the same NE are prevalent on different Web pages on a site, e.g., a person's name can be referred to in many ways. The proposed method groups similar NEs together in order to find these variants and align them by taking into account the string similarity of two NEs. Two NEs judged similar by their string similarity $StrSim(E1, E2)$ are more likely to be variants of the same NE if they appear on the same Web page or two Web pages which link to each other (we use the Levenshtein edit distance but other metrics are also suitable). The two NEs may appear on multiple Web pages, and we define the contextual distance $ConDis(E1, E2)$ between two NEs as the minimum number of links, regardless of link direction, between two Web pages where these two NEs appear. The contextual distance is zero if the two NEs both appear on the same Web page. We define the similarity between two NEs, *E1* and *E2*, as $Sim(E1, E2) = \frac{StrSim(E1,E2)}{a+b\times ConDis(E1,E2)}$, where a and b are weights.

2.3 Relation Strength and Ranking

For each target NE (which may be the person whose competencies we wish to discover), the relation strengths of co-occurring NEs are calculated by taking into account their number of co-occurrences with, and distances from the target. Associated NEs are then ranked and divided into separate lists for different types of NE (we used research areas, people, projects, and organizations). Thus NEs which have strong relations with the target can be identified. The relation strength between two NEs takes into account four aspects as follows.

Co-occurrence: Two NEs are considered to co-occur if they appear in the same Web page. Generally, if an NE is closely related to a target, they tend to co-occur more often. For two NEs, *E1* and *E2*, we use Resnik's method [8] to compute a relative frequency of co-occurrences of *E1* and *E2* as in Equation 1.

$$\hat{p}(E1, E2) = \frac{Num(E1, E2)}{N} \quad (1)$$

where $Num(E1,E2)$ is the number of pages in which $E1$ and $E2$ co-occur, and N is the total number of pages.

Distance: Two NEs which are closely related tend to occur close to each other. If two NEs, *E1* and *E2*, both occur only once in a Web page, the distance between them is the difference between their offsets. If *E1* occurs once and *E2* occurs multiple times in the Web page, the distance between *E1* and *E2* is the difference between the offset of *E1* and the offset of the closest occurrence of *E2*. When both *E1* and *E2* occur multiple times in the Web page, we average the distance from each occurrence of *E1* to *E2* and define the logarithm distance between *E1* and *E2* in the *i*th Web page as in Equation 2.

$$\overline{d_i}(E1,E2) = \frac{\sum_j (1+\log_2(\min(E1_j, E2)))}{Freq_i(E1)} \quad (2)$$

where $Freq_i(E1)$ is the number of occurrences of *E1* in the *i*th Web page and $\min(E1_j, E2)$ is the distance between the *j*th occurrence of *E1*, $E1_j$, and *E2*.

Frequency: An NE is considered to be more important if it has more occurrences in a Web page. Consequently, a numerous NE tends to have strong relations with other NEs which also occur on that page.

Page relevance: Given a target, *E1*, the weight of each Web page is given indicating its relevance in associating other NEs on the page with *E1*, e.g., for a person, a high relevance weight might be set to his/her homepage and a low relevance weight to his/her blog page.

Relation strength: Given a target, *E1*, we calculate the relation strength between *E1* and another NE, *E2*, by taking into account their co-occurrences, distance and frequency in co-occurred Web pages. The relation strength, $R(E1,E2)$, between *E1* and *E2* is defined in Equation 3.

$$R(E1,E2) = \hat{p}(E1,E2) \times \sum_i \left(\frac{w_i \times f(Freq_i(E1)) \times f(Freq_i(E2))}{\overline{d}_i(E1,E2)} \right) \quad (3)$$

where w_i is the weight showing the relevance of the *i*th Web page to *E1*, $f(Freq_i(E1)) = 1+\log_2(Freq_i(E1))$, $f(Freq_i(E2)) = 1+\log_2(Freq_i(E2))$, and $Freq_i(E1)$ and $Freq_i(E2)$ are the numbers of occurrences of *E1* and *E2* in the *i*th Web page respectively.

Thus the relation strength between a target and each of its co-occurring NEs is calculated. We rank co-occurring NEs in terms of their relation strengths with the target. Since these NEs are of different types, we divide the ranked list into a set of ranked lists for each type, e.g., lists of related people and related organizations.

We set a threshold, so that only relations with *R* above the threshold are selected. For example, we could set the threshold as the value at which two NEs co-occur with only one occurrence each, within a distance *D*, in only one Web page. Higher thresholds give high precision and low recall, and *vice versa*.

It is worth noting that, since the relation strength part of the algorithm comprises a combination of measures for co-occurrence, frequency and distance, the current algorithm has potential for refinement by substituting these components with others that are more sophisticated. Consider the case of judging the strength of the relation between two organizations *E1* and *E2*. If the algorithm were to be deployed in a semantic Web environment where the documents were annotated with reference to an ontology it would be possible to take account of instances below *E1* and *E2* in the taxonomic structure such as people employed by the organizations or subsidiary companies.

3. Approaches to Evaluation

The evaluation of the CORDER system presents problems typical of un-supervised machine learning in general when trying to establish if the algorithm has learnt a model that is fit for purpose. The main approaches to evaluation may be characterized as quantitative, gold standard and task oriented.

Quantitative methods judge whether the model produced is a "good" model based on quantifiable parameters. For example, a classic method for analyzing hierarchical agglomerative clustering is the cophenetic correlation coefficient [9], [10]. Square Error Criterion is commonly used to evaluate the efficiency of numerical data clustering [11]. Another method is Information Gain in which is possible to assess the quality of the clustering results over categorical data [12]. We are experimenting with this approach to evaluate a CORDER enhanced semantic clustering method (see Section 6).

Gold standard approaches compare the learned model to an "ideal" produced *a priori* by domain experts. These are typical in information retrieval and information extraction, e.g., the MUC series of competitions [13]. Their primary disadvantage is that standard collections are expensive to produce. Moreover, since they are based on expert opinion, they are intrinsically subjective.

Task oriented evaluations examine algorithms in the context of applications. They are concerned with whether the learning algorithm has produced a model that functions properly in use. Tonella et al. [14] discuss some of the problems associated with the task oriented approach including its cost and the need for careful design to minimize subjectivity.

Each approach has deficiencies. Therefore we favor a mixed strategy. Our first evaluation mined competencies using the website of the Knowledge Media Institute. This meant we had access to experts who could provide subjective data on the validity of the model. Current evaluation efforts include developing a quantitative method with which the rankings produced by CORDER can be benchmarked. In future, we hope to identify suitable gold standards and to evaluate CORDER on larger collections, e.g., the BBC news site. The CORDER system is currently being incorporated into a search system for finding subject experts in instant messaging groups (see section 6). Once this work is complete we can undertake a task-oriented evaluation.

4. Expert Evaluation

We created a web based form which allowed each expert to access the model that CORDER had generated for them. Thirteen people, representing a range of experience from PhD students to professors, modified their own model to produce rankings and relevance judgments closer to their own view of their interests and associations. These gave us a *post hoc* standard against which to measure CORDER's performance.

We used precision (P) and recall (R) to measure CORDER's ability to discover relevant NEs (eq. 4).

$$P_{T,User} = \frac{N_{User,CORDER,\mathrm{Relevant}}}{N_{CORDER,\mathrm{Relevant}}}$$
$$R_{T,User} = \frac{N_{User,CORDER,\mathrm{Relevant}}}{N_{User,\mathrm{Relevant}}} \quad (4)$$

where T is the type of NE, the number of NEs judged as relevant by CORDER is $N_{CORDER,\mathrm{Relevant}}$, the number of NEs judged as relevant by the user is $N_{User,\mathrm{Relevant}}$, and the number of NEs judged as relevant by both the user and CORDER is $N_{User,CORDER,\mathrm{Relevant}}$.

We used Spearman's coefficient of rank correlation RA [15] to measure CORDER's ability to rank NEs (eq. 5).

$$RA_{T,User} = 1 - \frac{6\sum_i (r_{i,User} - r_{i,CORDER})^2}{N_{User,CORDER,\mathrm{Relevant}}{}^3 - N_{User,CORDER,\mathrm{Relevant}}} \quad (5)$$

where $r_{i,User}$ and $r_{i,CORDER}$ $(1 \le i, r_{i,User}, r_{i,CORDER} \le N_{User,CORDER,\mathrm{Relevant}})$ are the two rankings provided by the user and CORDER respectively for the *i*th NE in the list. There are no ties in a set of rankings, i.e., for any two NEs, E_i and E_j $(i \ne j)$, $r_{i,User} \ne r_{j,User}$ and $r_{i,CORDER} \ne r_{j,CORDER}$. RA=1 when the two sets of rankings are in perfect agreement and RA_T=-1 when they are in perfect disagreement.

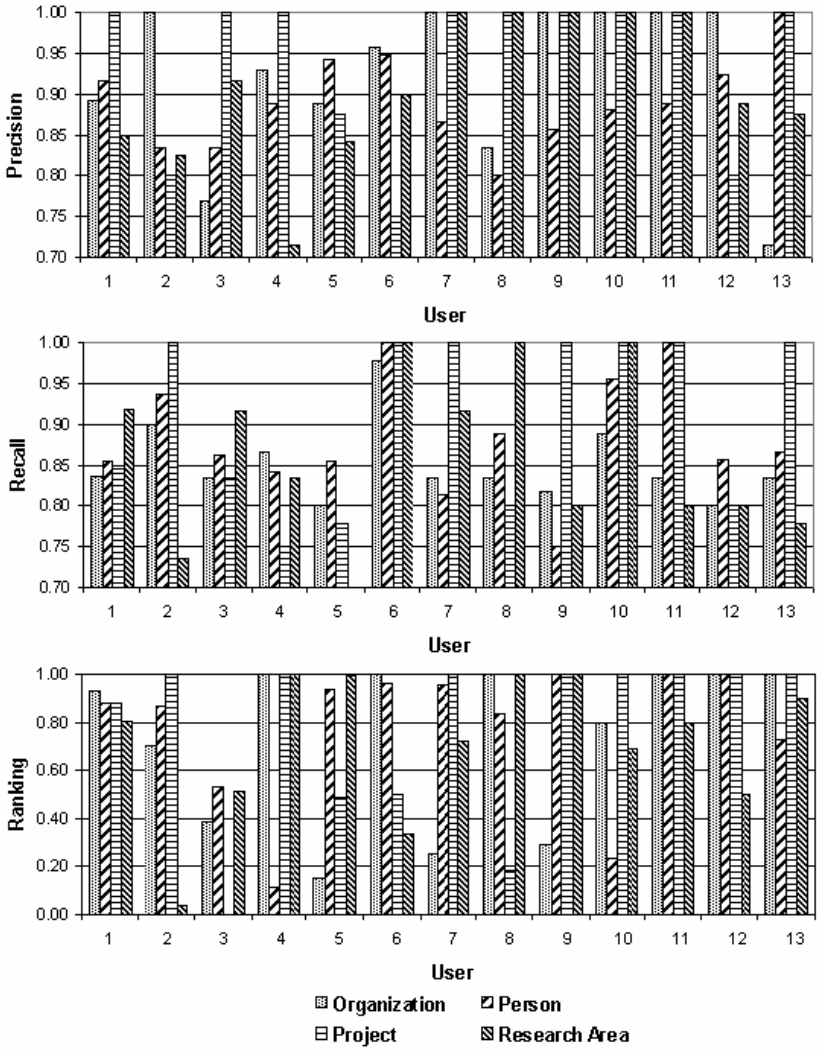

Figure 1: Precision, recall, and ranking accuracy of evaluation

The results (Figure 1) show that precision for all 13 users ranges between 70% and 100%. Recall ranges between 70% and 100%. RA ranges between 0 and 1.0. These figures suggest CORDER selects relevant NEs but doesn't always rank them the same as the experts.

The *post hoc* standard is imperfect in a number of ways which need to be addressed. The experts could only judge NEs that were found by CORDER. Some experts were inclined only to change the top of CORDER's rankings, i.e. the most relevant NEs. Some experts reported that it was hard to rank certain types of NEs, such as people, because their personal view of levels of importance was hard to quantify. Presenting experts with a randomized list to rank might give better results, but it would be a harder task. It may be that experts should instead be given a simpler task such as assigning NEs to groups such as "highly relevant", "relevant" and "not relevant".

5. Quantitative Benchmarking

As discussed in Section 3, we are following a mixed evaluation strategy. We are currently investigating a benchmarking method using the whole Web, represented by Google hits, as the standard. For each pair of named entities that CORDER judged to be relevant, Google is used to find the number of pages on which they co-occur. CORDER's ranking is then compared to a ranking based on the number of Google hits reusing the Spearman co-efficient (RA) described in Eq. 5. Thus the ranking that CORDER gets from in depth analysis of a representative subset of pages is benchmarked against a simple analysis of a larger number of pages.

Figure 2 compares user, CORDER and Google rankings for research area only. For the CORDER/Google comparison only three data points fall below 0.2 suggesting that CORDER's rankings are a reasonable model of the data found on the Web as a whole. Furthermore, CORDER got closer to the user rankings than Google (though it is important to remember that users started from CORDER rankings introducing a bias).

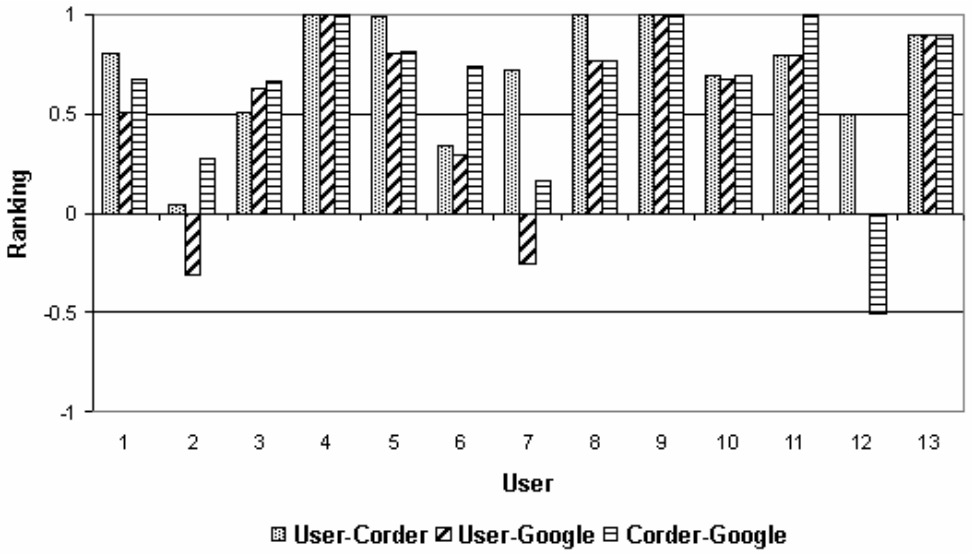

Figure 2. Ranking comparison for users, Google and CORDER

6. Conclusions and Continuing Work

We have shown that the CORDER algorithm can discover competency relationships that are judged to be appropriate by the people they concern. Our quantitative studies suggest that CORDER's rankings based on limited numbers of Web pages compare well to rankings based on hits on the whole Web.

This foundational work has encouraged us to start deploying the CORDER NE based ranking in a number of knowledge management scenarios. For example, it could be used on web portals to enhance the presentation of search results by presenting the documents most central for a topic or the best connected authors first. In addition, it could be used to mine text data for RDF triples to automatically input into a triple store.

The first competence discovery application we are building is the search service for the BuddySpace jabber environment (http://buddyspace.sourceforge.net/), called BuddyFinder. Finding useful contacts on instant messaging services is commonly based on registration information provided by the users. This has a number of weaknesses. In particular, users tend not to be motivated to provide more than a few keywords and the information can quickly go out of date. The BuddyFinder system asks them to supply the URL of their home page; users are more motivated to keep their home page comprehensive

and current than a profile on their instant messaging system. The CORDER algorithm uses the keywords in a query as the target NE and calculates the strength of the relations between the topic and users within the groups that the searcher belongs to based on data mined from their homepage and closely associated pages (for example blog pages). The results are presented as a ranked list of users (see Figure 3).

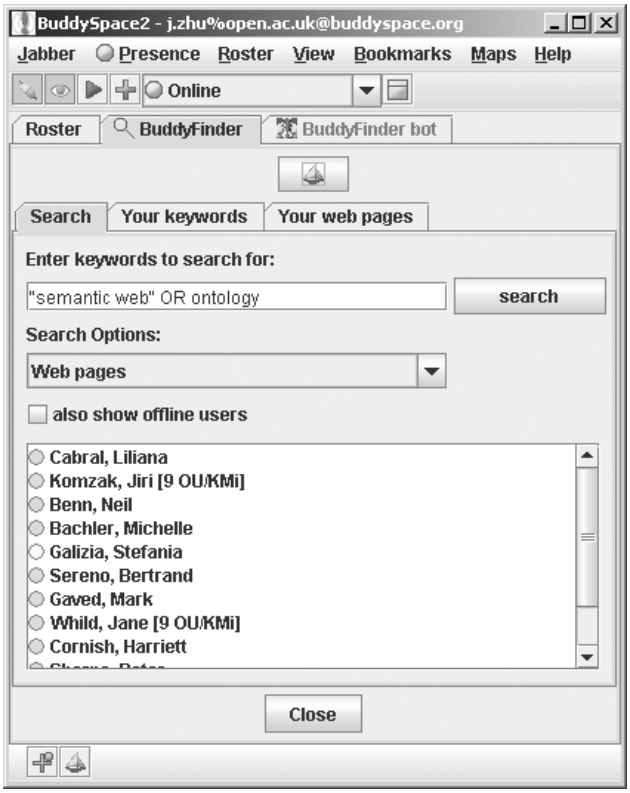

Figure 3. BuddyFinder output for a search on "semantic web" OR ontology

CORDER's rankings are derived from data mined from a collection of documents. In this way it gives a wider view of the "world" of a domain than data from a single document. We are experimenting with using the closest entities suggested by CORDER to improve the vector descriptions of documents for clustering. Our initial experiments suggest that this approach produces clusters which score as well as the widely used SOM method [16] on a total information gain measure of cluster quality. The execution time of the CORDER enhanced clustering method however increases linearly with the size and number of documents it examines so that it starts to outperform SOM on collections of more than 700 vectors. We intend to test this clustering approach on the Lattes Platform collection of curricula vitae discussed in Section 1.

Some refinements to the algorithm are required, including but not confine to:

- the introduction of a "timeline" to monitor changes in competencies; new ways to deal with noise and variants from the named entity recognizer,
- NLP methods to recognize the kind of relation indicated by a co-occurrence,
- sophisticated distance and relation strength metrics which exploit the power of ontologies (see Section 2 for discussion).

While there is still work to do we are optimistic that the CORDER algorithm is appropriate for use in competency discovery applications and has potential for application in other search scenarios where the ranking of entity data is desirable.

7. 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. This research was also partially supported by the Brazilian National Research Council (CNPq) with a doctoral scholarship held by Alexandre L. Gonçalves.

8. References

[1] C. Dozier, P. Jackson, X. Guo, M. Chaudhary, and Y. Arumainayagam, "Creation of an Expert Witness Database through Text Mining", In *Proceedings of the 9th international conference on Artificial intelligence and law*, 2003, pp. 177-184.

[2] O. Entzioni, "The World Wide Web: quagmire or gold mine?", *Communications of the ACM*, 1996, 39(11), pp. 65-68.

[3] S. Brin, "Extracting Patterns and Relations from World Wide Web", In *Proceedings of WebDB Workshop at 6th International Conference on Extending Database Technology*, 1998, pp. 172–183.

[4] E. Agichtein, and Gravano, "Snowball: Extracting Relations from Large Plain-Text Collections", In *Proceedings of the 5th ACM International Conference on Digital Libraries*, 2000, pp. 85–94.

[5] O. Etzioni, M. Cafarella, D. Downey, A. Popescu, T. Shaked, S. Soderland, S. Weld, and A Yates, "Methods for Domain-Independent Information Extraction from the Web: An Experimental Comparison", In *Proceedings of AAAI 2004*, 2004, pp. 391-398.

[6] T. Hasegawa, S. Sekine, and R. Grishman, “Discovering Relations among Named Entities from Large Corpora”, In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics*, 2004, 415-422.

[7] J. Zhu, V. Uren, and E. Motta, “ESpotter: Adaptive Named Entity Recognition for Web Browsing”, In *Proceedings of the 3rd Conference on Professional Knowledge Management (WM 2005)*, pp. 505-510.

[8] P. 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*, 1999, 11, pp. 95-130.

[9] R.R. Sokal and F.J. Rohlf, “The Comparison of Dendrograms by Objective Methods”, *TAXON*, 1962, 11, pp. 33-40, 1962

[10] M. Halkidi, Y. Batistakis and M. Vazirgiannis, “On Clustering Validation Techniques”, *Journal of Intelligent Information Systems*, 2001, 17(2/3), pp. 107-145.

[11] R. Duda, and P. Hart, *Pattern Classification and Scene Analysis*, Wiley, New York, 1973.

[12] C-H. Yun, K-T. Chuang, and M-S. Chen, Adherence Clustering: an Efficient Method for Mining Market-Basket Clusters”, *Information Systems*, In Press, 2005.

[13] DARPA (Defense Advanced Research Projects Agency), *Proc. of the Sixth Message Understanding Conference*, Morgan Kaufmann, 1995.

[14] P. Tonella, F. Ricca, E. Pianta, C. Girardi, G. Di Lucca, A. R. Fasolino, and P. Tramontana, “Evaluation Methods for Web Application Clustering**”,** In *Proceedings of the 5th International Workshop on Web Site Evolution***,** 2003.

[15] J.D. Gibbons, *Nonparametric Methods for Quantative Analysis*, Holt, Rinehart and Winston, 1976.

[16] T. Kohonen, *Self-Organizing Maps*, Springer Series in Information Sciences, Vol. 30, 1995.