Improving Web Data Annotations with Spreading Activation

Fatih Gelgi, Srinivas Vadrevu, and Hasan Davulcu

Department of Computer Science and Engineering,
Arizona State University,
Tempe, AZ, 85287, USA
{fagelgi, svadrevu, hdavulcu}@asu.edu

Abstract. The Web has established itself as the largest public data repository ever available. Even though the vast majority of information on the Web is formatted to be easily readable by the human eye, "meaningful information" is still largely inaccessible for the computer applications. In this paper, we present automated algorithms to gather meta-data and instance information by utilizing global regularities on the Web and incorporating with contextual information.
Experimental evaluations successfully performed on the TAP knowledge base and the faculty-course home pages of computer science departments containing 16,861 Web pages. The system achieves this performance without any domain specific engineering requirement.
Keywords: Semi-structured data, spreading activation, semantic partitioning.

1 Introduction

Scalable information retrieval [1] based search engine technologies have achieved wide spread adoption and commercial success towards enabling access to the Web. However, since they are based on an unstructured representation of the Web documents their performance in making sense of the available information is also limited.

Our system that we present in this paper is capable of gathering meta-data and instances from attribute rich Web page collections. Figure 1 displays an example of a faculty home page that lists the attribute information such as publications, address and telephone. Our system utilizes the presentation regularities within the Web page to produce an initial meta-tagging of the Web page using a semantic partitioning algorithm [2, 3], and utilizes (i) the global regularities on the Web and (ii) the contextual information to refine the meta-tagging using a *spreading activation network (SAN) framework*.

Our system differs from the previous work by not requiring template drivenness and an ontology. Thus, we can not readily use previously developed wrapper induction techniques [4–7] which require that the item pages should be template driven and ontology driven extraction techniques [8–10] which require that an

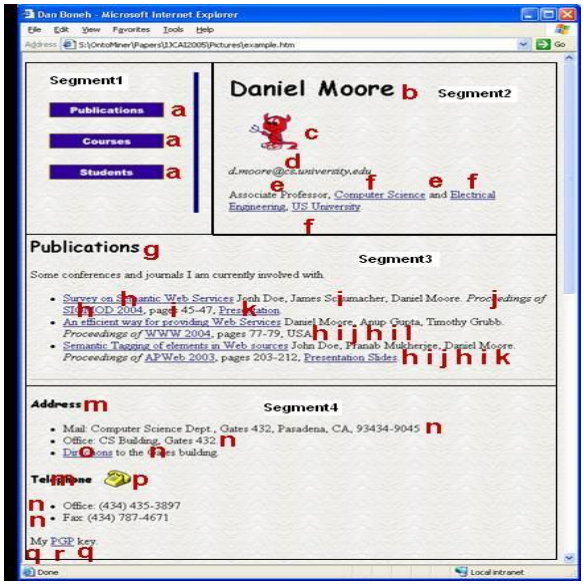

Fig. 1. An example of a Faculty home page. The labels in the page are marked with their corresponding path identifier symbols. The page segments in the Web page are also marked as Segments 1 to 4.

ontology of concepts, relationships and their value types is provided apriori in order to find matching information.

We will not discuss the details of our semantic partitioning algorithm in this paper, but more details about the algorithm can be found in [2, 3]. Initially our semantic partitioning algorithm also assigns ontological types to all labels based on their roles in the inferred group hierarchies using heuristic rules. The ontological type or the semantic role of a label can be either a concept (C), an attribute(A), a value (V) or nothing (N).

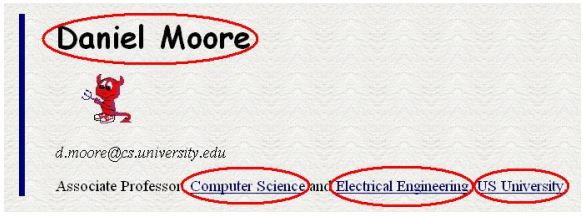

Fig. 2. The region from the facultys home in from the example in Figure 1. The circled labels are tagged incorrectly as attributes whereas they have to be values.

Presentation regularities on itself indeed won't be adequate for the semi-structured data such as web, thus the data generated by the semantic partitioning systems contain large amount of missing and incorrect annotations. For instance, in the example above *Daniel Moore, Electrical Engineering, Computer Science* and *US University* are all tagged as attributes instead of values as shown in Figure 2. The SAN framework is capable of reasoning with the meta-data regularities in order to improve the accuracy of the ontological types of labels found within object attribute pages.

The rest of the paper is organized as follows. Section 2 explains the overview of our system and Section 3 presents discusses our SAN framework in detail. Section 4 presents complexity analysis of our algorithms and Section 5 provides experimental results on various data beds. Finally we conclude in Section 6 with some future directions of our work.

2 System Overview

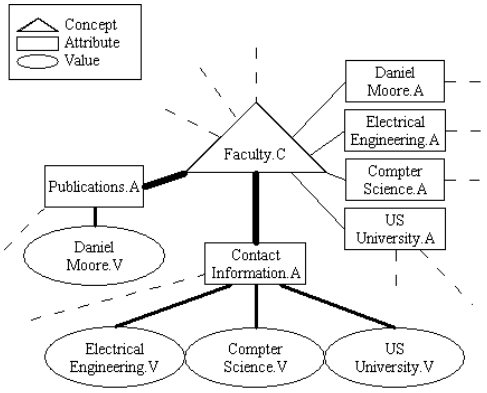

Fig. 3. A part of the graph of labels with roles and their association strengths are illustrated. Nodes with same roles have same shape. Each node has a label and its role which is concatenated at the end of the label (.C,.A,.V).

In order to distinguish various roles of labels, we need a model of their "context" that is able to accurately predict the role of a label. The context of a label and its role can be determined by its associations with other labels. Hence, a suitable model for reasoning with contexts is a relational graph structure where the nodes correspond to labels with roles and weighted edges correspond to association strengths between nodes. Figure 3 illustrates a relational structure of the tagged labels and their association strengths. *Daniel Moore, Electrical Engineering, Computer Science* and *US University* are all weakly associated with the *Faculty* concept as attributes. The same set of value tagged labels are also

strongly associated with the *Faculty* concept through *Publications* and *Contact Information* attributes.

In order to reason with the relational graph structures described above, we use the SAN [11, 12] which initializes tagged labels as nodes and their association strengths as weighted edges. Each label in the domain might have four nodes; label as a concept(C), an attribute(A), a value(V) or a nothing(N).

For each web page, SAN is activated with the nodes corresponding to the tagged labels of the page. Next, the energy spreads through the nodes proportional to their association strengths, which will also be propagated. The energy in the network will spread until it vanishes and then the activation levels of the nodes will be stabilized. Since the tagged labels are related to their associated neighbors as their context, it is expected that eventually the most activated role of each label will provide us with its context dependent correct role assignment.

For the above example, SAN was able to correct the roles of the mistagged attribute labels by iteratively spreading the energy through *Faculty.C* and then through *Publications.A* and *Contact Information.A*. Since the energy spreads proportional to the association strength eventually *Daniel Moore.V*, *Electrical Engineering.V, Computer Science.V* and *US University.V* will be activated more than their attribute nodes thus yielding the correct annotations.

The next section, formalizes the above SAN in details.

3 Technical Details

SAN works in three phases: *initialization*, *activation* and *inference*. The first phase is the construction of the relation graph and the remaining phases are for querying the graph. A *query* is the set of tagged labels in the web page. In the activation phase, the energies of the corresponding nodes in the query are initialized and SAN runs. The last phase is for collecting the roles of the labels in the query assigned by SAN.

3.1 Initialization

In the initialization phase the nodes and their association strengths are initialized. For the rest of the paper, we will use the formalism, $\mathcal{L}$ for the set of the labels in the domain, and l_i and r_i to express the label and role of a node i where $l_i \in \mathcal{L}$ and $r_i \in \{C, A, V, N\}$. In the network, the association strength w_{ij} between two nodes i and j is initialized using the following equation,

$$w_{ij} = w_{ji} = \frac{|l_i.r_i \leftrightarrow l_j.r_j|}{|l_i.r_i| + |l_j.r_j|}, \tag{1}$$

where $|l_i.r_i|$ specifies the frequency of label l_i associated with the role r_i in the data, and $|l_i.r_i \leftrightarrow l_j.r_j|$ specifies the frequency of edges between $l_i.r_i$ and $l_j.r_j$ in the semantic hierarchies.

Although meta-data structures are hierarchical, i.e. with directed edges, we used undirected edges in our SAN model in order to enable energy spread to the

neighboring context nodes as an average of their directed association strengths. Otherwise, for instance the attributes in the context wouldn't activate their corresponding concept. On the other hand, "being related" does not mean the hierarchical structure, but bi-directional relationship.

3.2 Activation

Each node i in SAN has an *activation level*, A_i and a *spreading energy*, S_i. In the activation phase, we use an iterative approach. For each Web page, we set the initial spreading energies of the nodes corresponding to the tagged labels of the page with their support. Then in each iteration the spreading energies spread to the neighbors with a decay factor α, which adjusts the total energy of the network to reduce in each iteration. Thus, it is guaranteed that the algorithm stops after a number of iterations, i.e., when the total spreading energy of the network becomes negligible. Initial values and recurrence formulas of A_i and S_i are given in the following:

$$\begin{aligned} A_i^{(0)} &= 0 \\ A_i^{(t+1)} &= A_i^{(t)} + \sigma \sum\nolimits_j \frac{w_{ij}.S_j^{(t)}}{\sum_k w_{jk}} \end{aligned} \tag{2}$$

for all nodes i in SAN and

$$\begin{aligned} S_i^{(0)} &= \begin{cases} \frac{|l_i.r_i|}{|\mathcal{D}|} & \text{, if } l_i.r_i \in \mathcal{Q} \\ 0 & \text{, otherwise} \end{cases} \\ S_i^{(t+1)} &= \alpha\sigma \sum\nolimits_j \frac{w_{ij}.S_j^{(t)}}{\sum_k w_{jk}} \end{aligned} \tag{3}$$

where $|\mathcal{D}|$ is the number of Web pages in the domain, $\mathcal{Q}$ is the current query and iteration numbers are denoted by superscripts. σ is a normalization factor such as $\frac{1}{\sum_i \sum_j w_{ij}}$ or $\frac{1}{\max_i \{\sum_j w_{ij}\}}$. And, at a particular iteration t, the termination criterion for the activation is,

$$\sum\nolimits_i S_i^{(t)} < \beta \tag{4}$$

where β is a very small number.

3.3 Inference

In the inference phase, for each label of the Web page we determine the node with the highest activation levels to reassign its role. Formally, the role of a label l is,

$$r_{\arg\max_i \{A_i | l_i = l\}}. \tag{5}$$

4 Complexity Analysis

In this section, we will show that the above SAN can be very efficiently implemented thus yielding a fast and scalable model.

Assuming there are n labels and m associations between labels, in the initialization phase requires only $O(n + m)$ time. For the rest of the analysis we will assume that we are working on a very large $n >> m$ which is reasonable for the web data. Hence, this phase has $O(n)$ time and memory complexity.

In the activation phase, suppose we have p web pages. Another issue is the number of iterations required per page. In each iteration, since the total activation decreases by a constant factor, α, then the number of iterations, k, will also be a constant. Finally, since SAN runs for each web page, the time complexity for all activations will be $O(pn)$.

5 Experimental Results

In this section we provide the details about the test beds we used in our experiments and present the results with these data sets. Precisions, recalls and f-measures given in the results are calculated regarding to the number of correctly and incorrectly tagged labels in the data. Experimental data and evaluation results are also available online at `http://cips.eas.asu.edu/ontominer_files/eval.htm`.

5.1 Setup

TAP Dataset: The TAP data set contains selected categories from TAP Knowledge Base 2 (TAP KB2) [13], including AirportCodes, CIA, FasMilitary, GreatBuildings, IMDB, MissileThreat, RCDB, TowerRecords and WHO. These categories alone comprise $9,068$ individual Web pages and these Web pages are attribute-rich. We provide experimental results for this data set with our algorithms and compare them against the relations obtained by TAP.

CSEDepts Dataset: The CSEDepts data set consists of 60 computer science department Web sites, comprising $7,793$ individual Web pages and $27,694$ total number of labels. The computer science department Web sites are *meta-data-driven*, i.e., they present similar meta-data information across different departments. To demonstrate the performance of our meta-tagging algorithms, we created a smaller data set containing randomly chosen 120 Web pages from each of the *faculty* and *course* categories. We provide experimental results for this data set with our algorithms.

5.2 Preprocessing TAP Dataset

To test the SAN framework, we converted RDF files in the TAP knowledge base into triples then we applied distortions to obtain SAN inputs. SAN needs three files as inputs: (1) counts of tagged labels in the overall data, (2)

tagged label pairs which are related (for the relational graph of SAN) such as $(concept, attribute)$ or $(attribute, value)$, and (3) sets of tagged labels for Web pages.

Regular Expression Processing: During the conversion we first split values into tokens and preprocessed the common types of values in the triples such as percentage, month, real number, currency etc... using simple regular expressions.

Distortion: For synthetic data, we considered real world situations and tried to prepare the input data as similar as possible to the data on the Web. There are two types of distortion: *deletion* and *role change*. Setting the distortion percentages for both deletion and role change first, we used the percentages as distortion probabilities for each tagged label in the Web page in our random distortion program.

Over TAP data set, we prepared test cases for three kinds of distortions. In the first one, we only applied deletion with different percentages. In the second, similarly we only applied role changing. And the last one is the mixture of the previous two; we applied the same amount of deletions and role changes.

Web sites	# of Web pages	Average # of labels per page	Distortion rates (role changes)											
			5%			20%			40%			60%		
			P	R	F	P	R	F	P	R	F	P	R	F
AirportCodes	3829	34	0.96	0.90	0.93	0.96	0.90	0.93	0.96	0.88	0.92	0.95	0.86	0.90
CIA	28	1417	0.96	0.85	0.90	0.92	0.82	0.87	0.81	0.74	0.77	0.47	0.43	0.45
FasMilitary	362	89	0.88	0.71	0.78	0.82	0.65	0.73	0.70	0.55	0.62	0.58	0.41	0.48
GreatBuildings	799	37	0.92	0.77	0.84	0.91	0.75	0.82	0.87	0.71	0.78	0.81	0.64	0.71
IMDB	1750	47	0.91	0.79	0.85	0.89	0.76	0.82	0.87	0.72	0.79	0.83	0.65	0.73
MissileThreat	137	40	0.99	0.96	0.98	0.98	0.94	0.96	0.97	0.92	0.94	0.81	0.77	0.79
RCDB	1637	49	0.98	0.91	0.94	0.95	0.88	0.92	0.91	0.85	0.88	0.84	0.77	0.80
TowerRecords	401	63	0.23	0.83	0.77	0.88	0.80	0.84	0.88	0.77	0.82	0.86	0.72	0.78
WHO	125	21	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Overall	9068	200	0.87	0.86	0.89	0.92	0.83	0.88	0.89	0.79	0.84	0.79	0.69	0.74

Table 1. SAN performance for individual Web sites. P, R and F stands for precision, recall and f-measure respectively.

5.3 Experiments with the TAP Data Set

We present our results in two categories: (1) individual Web sites and (2) mixture Web sites. For both categories, SAN performed with 100% accuracy for distorted data with only deletions. The reason is, deletion does not generate ambiguity since the initial data is unambiguous. Thus, we found unnecessary to include them in the tables and figures.

Web sites	Distortion rates (role changes)			
	5%	20%	40%	60%
AirportCodes	1.00	0.98	0.95	0.92
CIA	0.99	0.96	0.89	0.66
FasMilitary	0.99	0.93	0.82	0.65
GreatBuildings	0.99	0.95	0.88	0.78
IMDB	0.99	0.95	0.89	0.79
MissileThreat	1.00	0.99	0.97	0.87
RCDB	1.00	0.98	0.94	0.86
TowerRecords	0.99	0.96	0.91	0.83
WHO	1.00	1.00	1.00	1.00
Average	0.99	0.97	0.92	0.82

Table 2. Label accuracies for individual Web sites.

Experiments with the individual Web sites provided us encouraging results to start experiments with mixture of Web sites as shown in that Tables 1 and 2. Table 1 displays SAN's performance of annotation on distorted data and Table 2 displays the final accuracies of Web sites which are initially distorted with {5,20,40,60} percent role changes. Overall results show that even for 60% role changes SAN performed with 82% accuracy. The performance is usually better with the Web sites containing large number of Web pages due to the high consistency and regularity among the Web pages. Another factor is the size of the tagged label set in the Web pages. The larger the set, the more difficult to keep context concentrated on the related roles in ambiguous data. That played the most important role for the low performance with *CIA* and *FasMilitary* Web sites and, high performance with *WHO* and *AirportCodes*.

In the experiments with mixture of Web sites, we tested SAN with {5,20,40,60,80} percent role changes and {(5,5),(15,15),(25,25),(35,35),(45,45)} percent (deletion, role change) distortions. Figures 4 and 5 show the performance of SAN in terms of concept, attribute and value accuracy, and the final Web site accuracy respectively. The overall performance for the mixture Web sites is slightly lower than the individual Web sites due to the fact that mixture Web sites initially have some ambiguities.

In conclusion, the overall results show that SAN can recover the TAP data up to 75% even with 60% and (35%,35%) distortion. the results are not surprising since the data set is template driven and also the relations are not complicated. Verifying the robustness of the system with template driven Web sites, next we will give the experimental results with a non-template driven data set.

5.4 Experiments with the CSEDepts Data Set

Recalling the motivating example, one can observe that labels in the faculty and course home pages of computer science departments are highly ambiguous. Many labels have different roles in different Web pages depending on the context.

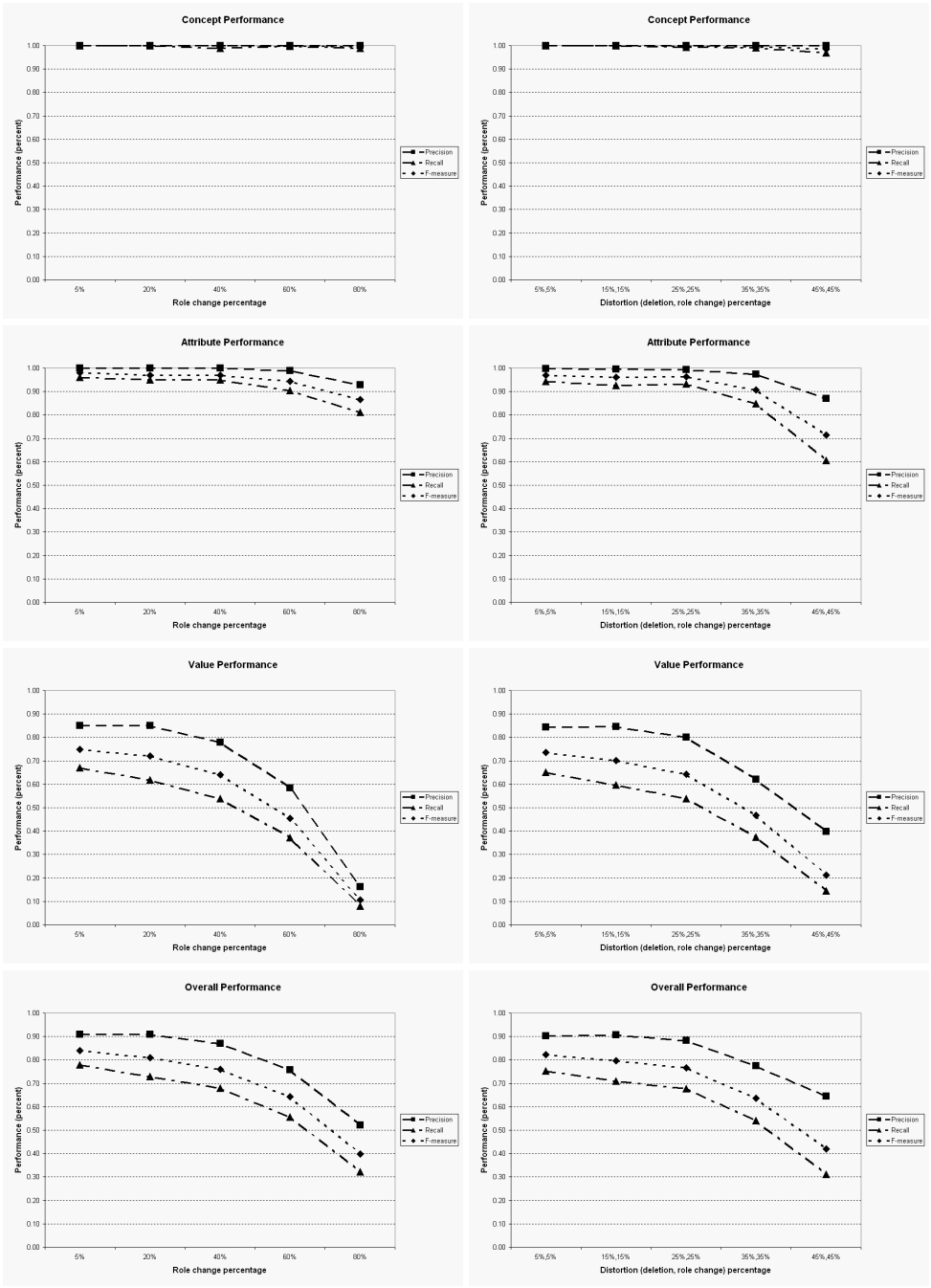

Fig. 4. SAN performance on mixed data for various distortion rates. In the left column, distortions are only role changes whereas in the right column, distortions include both deletions and role changes.

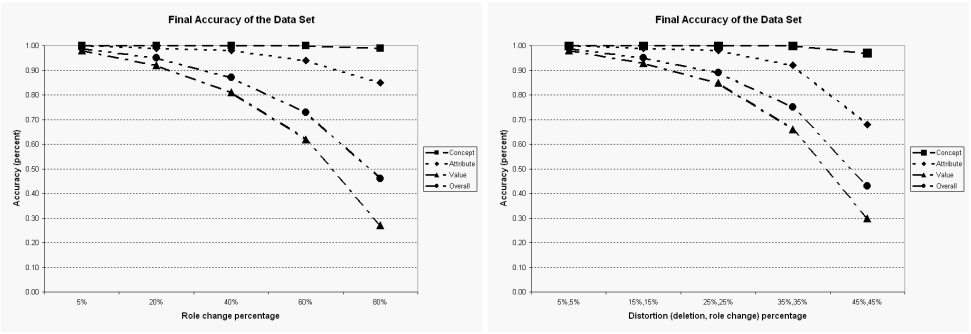

Fig. 5. Final label accuracies for different distortions. In the left figure, distortions are only role changes whereas in the right figure, distortions include both deletions and role changes.

Domain	Method	Total Acc.	P(C)	R(C)	F(C)	P(A)	R(A)	F(A)	P(V)	R(V)	F(V)	P(N)	R(N)	F(N)
	Sem.Part.	57	0	0	0	64	42	50	61	75	76	21	90	37
Courses	Majority	63	50	0	7	69	48	55	72	83	77	31	85	48
	SAN	81	83	94	88	82	86	83	91	80	85	44	85	62
	Sem.Part.	62	50	0	4	51	44	47	65	80	71	34	31	32
Faculty	Majority	62	7	0	2	62	51	56	64	85	73	56	35	42
	SAN	83	86	95	90	74	83	79	93	79	86	60	84	69
	Sem.Part.	60	25	0	2	57	43	49	70	78	73	27	61	35
Overall	Majority	65	28	0	5	66	48	55	68	64	75	43	60	45
	SAN	82	85	95	89	80	84	81	92	80	85	52	84	65

Table 3. Comparison of annotations of the data using the initial semantic partitioner based method, the baseline majority method, and the spreading activation based method. P(C), R(C), F(C) denote the precision, recall and f-measure of the concept annotation. Similar notation is used for attributes (A), values (V) and nothing (N).

In this experiment, we used the semantic partitioning algorithm to obtain initial annotations of the labels from Web pages. The initial annotations have low accuracy which is even eligible to provide enough statistics for SAN.

To compare our result, we used a simple but straightforward idea, named *majority measure*, to infer the role for each label by simply associating its majority role within the domain. The underlying assumption is that each label has a "unique role" in the domain. Although, this simple method improves the annotation accuracy slightly, the majority of the incorrect meta-data tags still remains weakly annotated as shown in the experimental results in Figure 6 and Table 3. Besides, we observed that the "unique role assumption" did not hold in many cases since labels might have different roles in different contexts. Figure 2 also demonstrates an example of such a case.

The individual f-measure values as shown in Figure 6 demonstrate that spreading activation based annotation achieves better performance in all the four different role annotations. The number of concept labels in the domain are fewer than other ontological types. Thus the lack of sufficient statistics for the concept annotation result in a surge in the increase of the f-measure for spreading activation, i.e., even though spreading activation corrects a few concept annotations, it is boosted because there are only a few number of concept labels. The f-measure for value annotation is the highest among all role annotations, as the semantic partitioner is able to annotate the values correctly in most cases and spreading activation is able to correct the remaining annotations.

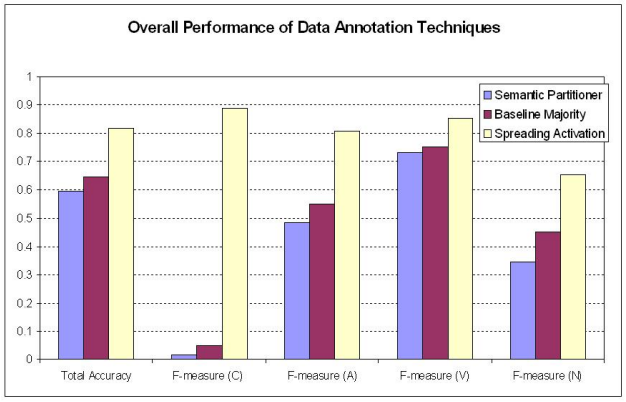

Fig. 6. Comparison of f-measure values for various data annotation techniques.

These experimental results are obtained by comparing the data annotations of the algorithms to manually annotated data by eight human volunteers. The inter-human agreement on manual annotation was 87%, which indicates that the data annotations can be ambiguous and can be interpreted differently in various contexts. But our algorithms are able to perform well even in the presence of such ambiguity.

6 Conclusions & Future Work

In this paper, we have presented a system that can automatically gather and separate meta-data and their instances from various kinds of Web pages. The main focus has been on the SAN framework which is capable of reasoning on semi-structured data obtained from our semantic partitioning system. The novelty of the framework comes from spreading the energy of the contextual information thorough the overall relational graph to improve the annotations. That makes the system utilize the global regularities incorporating the contextual information.

Many research questions remain open for future work. We could identify the missing attribute labels of the values in the Web pages by identifying them using the contextual information in the spreading activation network. We could use the final annotations provided by the meta-tagging algorithm to bootstrap the semantic roles for the labels in the semantic partitioning algorithm, so it can perform the hierarchical grouping algorithm much more efficiently.

References

1. Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.
2. Hasan Davulcu, Srinivas Vadrevu, Saravanakumar Nagarajan, and I.V. Ramakrishnan. Ontominer: Bootstrapping and populating ontologies from domain specific web sites. *IEEE Intelligent Systems*, 18(5), September 2003.
3. Srinivas Vadrevu, Saravanakumar Nagarajan, Fatih Gelgi, and Hasan Davulcu. Automated metadata and instance extraction from news web sites. In *The 2005 IEEE/WIC/ACM International Conference on Web Intelligence*. Compiegne University of Technology, France, 2005 (to appear).
4. Naveen Ashish and Craig A. Knoblock. Semi-automatic wrapper generation for internet information sources. In *Conference on Cooperative Information Systems*, pages 160–169, 1997.
5. Nickolas Kushmerick, Daniel S. Weld, and Robert B. Doorenbos. Wrapper induction for information extraction. In *Intl. Joint Conference on Artificial Intelligence*, pages 729–737, 1997.
6. Valter Crescenzi, Giansalvatore Mecca, and Paolo Merialdo. Roadrunner: Towards automatic data extraction from large web sites. In *Proceedings of 27th International Conference on Very Large Data Bases*, pages 109–118, 2001.
7. Arvind Arasu and Hector Garcia-Molina. Extracting structured data from web pages. In *ACM SIGMOD*, San Diego, USA, 2003.
8. Oren Etzioni, Michael Cafarella, Doug Downey, Stanley Kok, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. Web-scale information extraction in knowitall. In *Intl. World Wide Web Conf.*, 2004.
9. Fabio Ciravegna, Sam Chapman, Alexiei Dingli, and Yorick Wilks. Learning to harvest information for the semantic web. In *Proceedings of the 1st European Semantic Web Symposium*, Heraklion, Greece, 2004.
10. Stephen Dill, John A. Tomlin, Jason Y. Zien, Nadav Eiron, David Gibson, Daniel Gruhl, R. Guha, Anant Jhingran, Tapas Kanungo, Sridhar Rajagopalan, and Andrew Tomkins. Semtag and seeker: Bootstrapping the semantic web via automated semantic annotation. In *Twelth International Conference on World Wide Web*, pages 178–186, 2003.
11. A.M. Collins and E.F. Loftus. A spreading activation theory of semantic processing. *Psychological Review*, (82):407–428, 1975.
12. G. Salton and C. Buckley. On the use of spreading activation methods in automatic information. In *SIGIR '88: Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 147–160. ACM Press, 1988.
13. R.V. Guha and Rob McCool. Tap: A semantic web toolkit. *Semantic Web Journal*, 2003.