Ontology-Driven Semantic Matches between Database Schemas

Sangsoo Sung and Dennis McLeod
Department of Computer Science
and Integrated Media System Center
University of Southern California,
Los Angeles, California, 90089-0781
{sangsung, mcleod}@usc.edu

Abstract

Schema matching has been historically difficult to automate. Most previous studies have tried to find matches by exploiting information on schema and data instances. However, schema and data instances cannot fully capture the semantic information of the databases. Therefore, some attributes can be matched to improper attributes. To address this problem, we propose a sche- ma matching framework that supports identification of the correct matches by extracting the semantics from ontologies. In ontologies, two concepts share similar semantics in their common parent. In addition, the parent can be further used to quantify a similarity bet- ween them. By combining this idea with effective cont- emporary mapping algorithms, we perform an ontolo- gy-driven semantic matching in multiple data sources. Experimental results indicate that the proposed method successfully identifies higher accurate matches than those of previous works.

1. Introduction

With the steady advancement of semantically rich data representations in database systems, similar domains have been illustrated in different manners and in diverse terminologies by domain experts who typically have their own interpretations and analysis of the domain. Integrating data from heterogeneous databases thus yields many new information management challenges. In particular, schema matching is inherently difficult to automate and has been regarded as a tedious and error-prone task since schemas typically contain limited information without semantics of attributes.

In most previous studies, schema matching has been in general performed by gathering information for mapping from various phases of an attribute including its name, type, patterns and statistics of data instances [1-6]. For example, by comparing names, types, and sample instances between attributes "phone number" and "telephone" in compatible tables, these two attributes can be matched. However, schema and data instances thus cannot fully capture the meanings. If we only consider patterns of instances, domain, and name of attributes "phone number" and "fax number", these two names can then be matched. Therefore, excluding semantic information of the attributes is limited to discovering appropriate matches between database schemas. By illuminating the difficulties posed by lack of semantics, we have shown that there is a need for an alternative method to obtain semantics of data from external data sources. Toward this end, our approach is to incorporate ontologies to gain semantic information of data.

The goal of this paper is to introduce, define, and quantify mapping frameworks that support mechanic-sms for interconnecting similar domain schemas. We divide the mapping algorithms into two categories: the semantics-driven mapping framework utilizes the semantics of data, which is captured by ontologies, while the data-driven mapping framework also depends on efficient matching algorithms, which have been introduced in the previous research.

In order to achieve this goal, we hypothesize that ontology-driven schema matching can improve matching accuracy, since it can support the capture of suffici-ent semantic information of data while the traditional methods cannot. To evaluate this hypothesis, we 1) define a semantics-driven mapping framework and a data-driven mapping framework, 2) quantify the degree of similarity using ontologies and schema information, and 3) combine the similarities which are produced by both mapping frameworks.

The remainder of this paper is organized as follows: In Section II, we illustrate both mapping frameworks.

Section III discusses the experimental results. Finally, Section IV concludes this study.

2. Schema Matching

To discover the correspondences in the schema *S* and *T*, we compute similarity matrix M_{ST} for *S* and *T*. *S* has attributes s_1, s_2, … , s_n, and *T* has attributes t_1, t_2, … , t_n. M_{ST} is a matrix

$$\begin{bmatrix} SIM(s_1,t_1) & \cdots & SIM(s_1,t_m) \\ \vdots & SIM(s_i,t_j) & \vdots \\ SIM(s_n,t_1) & \cdots & SIM(s_n,t_m) \end{bmatrix},$$

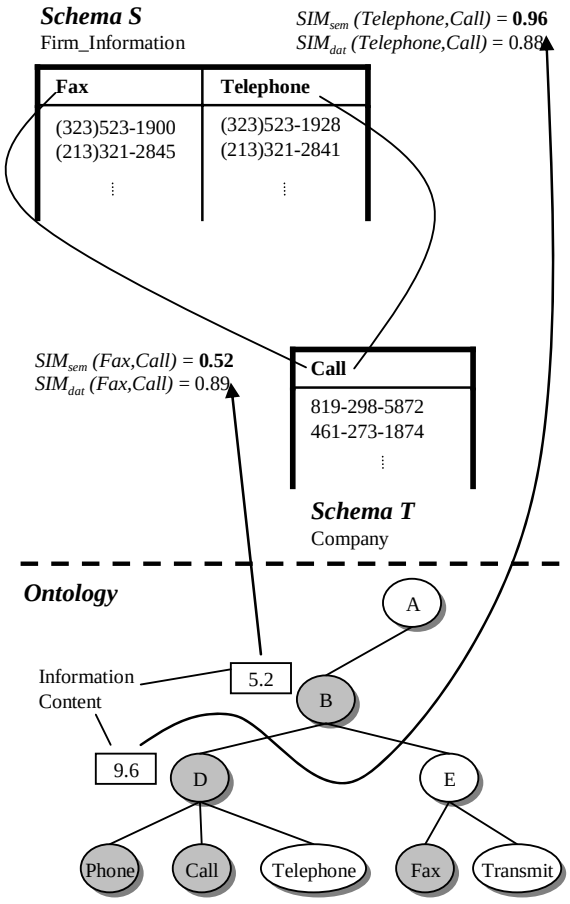
where $SIM(s_i, t_j)$ is an estimated similarity between the attributes s_i and t_j $(1 \le i \le n, 1 \le j \le m)$.

To find the most similar attribute in the other schema, we propose an ontology-driven mapping algorithm with an ensemble of multiple matching methods. The mapping algorithms are mainly divided into a semantics-driven mapping framework and a data-driven mapping framework. The former generates the matches based on information content [7], while the latter performs the matches based on the premise that the data instances of similar attributes are typically co- ngruent. Both frameworks thus increase the accuracy of similarity by mutual complementation. Each framework produces a mapping matrix; respectively M_{ST}^{sem} and M_{ST}^{dat}. Thus, the similarity matrix M_{ST} is

$$M_{ST} = \alpha \cdot M_{ST}^{sem} + \beta \cdot M_{ST}^{dat}, \; where \; \alpha + \beta = 1$$

Leveraging a mapping based on the meaning of the attributes achieves a level of matching performance that is significantly better than a conventional schema matcher. Two techniques contribute to the matching accuracy:

- **Matching ambiguity resolution**:
  It can identify actual mappings although they are ambiguous.
- **Providing candidates that refer to a similar or same object**:
  It also provides matching candidates even if the data-driven framework fails to select the candidates.



SIM_{sem} is an estimated similarity using the semantics-driven mapping
SIM_{dat} is an estimated similarity using the data-driven mapping

Figure 1. Matching ambiguity can be resolved with semantics-driven mapping framework

In our example, the values of the instances of the attributes "Fax" and "Telephone" in the table "Firm_Information" of schema S and the attribute "Call" in the table "Company" of schema T share common patterns. As shown in **Figure 1**, the estimated similarities resulting from the data-driven mapping framework are too close to determine which correspondence is more suitable for this mapping. However, the semantics-driven mapping framework provides increased evidence for the mapping between the attributes "Telephone" and "Call" since both words are semantically more related than the attribute pair of "Fax" and "Call". Therefore, it is necessary to prune candidate mappings.

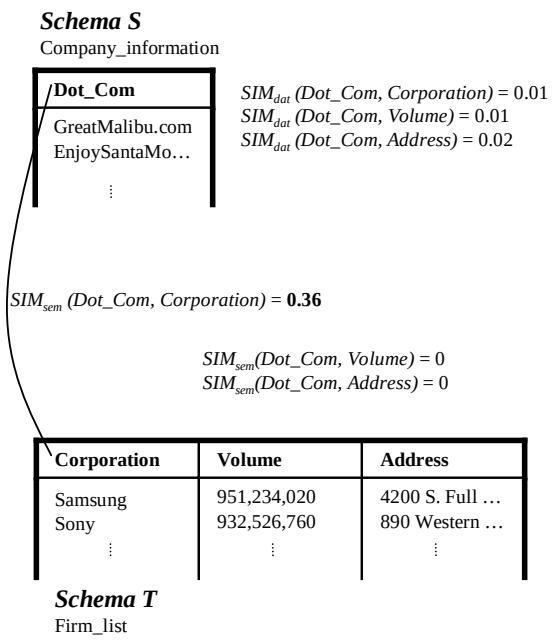


Figure 2. The semantics-driven framework provides candidate mappings

In the other example, the values of the instances of the attribute "Dot_Com" of the table "Firm_list" in schema S are dissimilar to all the attributes of the table "Company_information" in schema T. **Figure 2** illustrates that the data-driven framework fails to find the mapping candidates. However, the correspondence between "Corporation" and "Dot_Com" can be found by the semantics-driven mapping framework because the value of their information content is relatively higher than the other mappings.

2.1. Data-Driven Mapping Framework

The attribute names in the schemas can be very difficult to understand or interpret. In this section, we propose a framework that functions correctly even in the presence of opaque attribute names using the data values. Previous research has shown that an effective matching technique utilizes an ensemble of searching overlap in the selection of the data types and represent- ation of the data values, comparing patterns of the data instances between the schema attributes, linguistic mat- ching of names of schema attribute, and using learning techniques [1, 3, 4, 6]. In the data-driven mapping framework, we mainly make use of the fact that the schemas, which we are matching, are associated with the data instances we have.

By comparing the attribute instances, the mapping can be found since the similar attributes share similar patterns or representations of the data values of their instances. Thus, there are two types of base matchers such as the pattern-based matcher and the attribute-based matcher.

2.1.1. Pattern-based Matcher

The pattern-based matcher tries to find a common pattern of the instance values, such as fax/phone numb- ers, or monetary units. It determines a sequence of alp- habets, symbols and numbers that are most character- istic in the instances of an attribute. Given any value of the instances, we transform each alphabet to "A", sym- bol to "S", and number to "N". To compute the simil- arity, it compares the patterns by calculating the values of the edit distance [8] of a pair of patterns. An edit di- stance between two strings is given by the minimum number of the operations needed to transform one string into the other, where an operation can be either an insertion, deletion, or substitution. For example, "(213)321-4321" is transformed into "SNNN- SNNNSNNNN" and "213-321-4321" is transformed into "NNNSNNNSNNNN". In this case, the edit distance between two numbers is 1.

Let a_i and b_j be instances of the attribute *s* and *t* $(1 \le i \le N_a, 1 \le j \le N_b)$. Let $EditDist(a_i, b_j)$ denote an edit distance value between the patterns of the attribute instances a_i and b_j. In addition, it also contributes to a performance to use top a_i most frequent instances because pairwise comparison is typically a time consuming task. Let g_i denote the number of the instance a_i in the attribute *s*, and h_j be the number of the instance b_j in the attribute *t*. We assume that a_i and b_j are sorted in a descending order with respect to g_i and h_j. The similarity between the instance patterns of the attributes s_i and t_j can be quantified as follows:

$$SIM_{pat}(s,t) = \sum_{i=j=1}^{k} \left\{ \frac{1}{2}\left(\frac{g_i}{N_a} + \frac{h_i}{N_b} \right) \times \frac{1}{EditDist(a_i, b_j) + 1} \right\}$$

By detecting the most *k* frequent pattern, we can use the pattern to find a match with the pattern of the corresponding attributes.

2.1.2. Attribute-based Matcher

The attribute-based matcher tries to find common properties of the attributes. Comparing various phases of the attributes such as name and domain information also provides the correspondence between the attributes [1, 3-6]. Thus, the attribute-based matcher maps attributes by comparing the attribute's names and types.

Comparison of the names among the attributes is performed only when the domain information of two attributes is similar. Due to a number of diverse ways to represent the names of the attributes like compound words, we compute a prediction based on the frequency of the co-occurred N-gram of the attributes' name. Tri-gram was the best performer in our empirical evaluation.

Let $SIM_{dat}^{atr}(s_i, t_j)$ be a prediction, which is produced by this attribute-based matcher. Thus, the similarity from the data-driven mapping framework can be defined as:

$$SIM_{dat}(s_i, t_j) =$$
$$\alpha \cdot SIM_{dat}^{pat}(s_i, t_j) + \beta \cdot SIM_{dat}^{atr}(s_i, t_j)$$
$$where\ \alpha + \beta = 1$$

Unfortunately, this mapping framework is not always successful as indicated in **Figure 1** and **Figure2**. When it fails to find mappings, it is often because of its inability to incorporate the real semantics of the attributes to identify the correspondences. In the following sections, we propose a technique to resolve this problem.

2.2. Semantics-Driven Mapping Framework

The semantics-driven mapping framework tries to identify the most similar semantics of attribute in the other schema when the attribute names are not opaque. The name of an attribute typically consists of a word or compound words that contains the semantics of the attribute. Thus, the semantic similarity between s_i and t_j can be measured by finding how many words in two attributes are semantically alike. We describe how we measure a semantic similarity.

2.2.1. Semantic Similarity

Previous research has measured semantic similarity, which is based on statistical/topological information of the words and their interrelationships [9]. An alternative approach has recently been proposed to evaluate the semantic similarity in a taxonomy based on information content [7, 10]. Information content is a corpus-based measure of the specificity of a concept. This approach relies on the incorporation of the empirical probability, which estimates into a taxonomic structure. Previous research has shown that this type of approach may be significantly less sensitive to link density variability [9, 10]

Measures of the semantic similarity in this approach quantify the relatedness of two words, based on the information contained in an ontological hierarchy. Ontology is a collection of the concepts and inter-relationships [11]. There are two types of inter-relationships: a child concept may be an instance of its parent concept (is-a relationship), or a component of its parent concept (part-of relationship). In addition, the child concept can have multiple parents, thus there may exist multiple paths between the child concept and the parent concept. WordNet, which is a lexical database, is particularly well suited for similarity measures, since it organizes nouns and verbs into hierarchies of *is-a* or *part-of* relations. Thus, we have employed WordNe- t::Similarity [12], which has implemented the semantic relatedness measures that compute information content using WordNet ontology from untagged corpora such as the Brown Corpus, the Penn Treebank, and the British National Corpus [12].

Let c denote a word of an attribute. The information content of a word *w* can be quantified as follows:

$$IC(w) = -\log(p(w))$$

where $p(w)$ is the probability of how much word w occurs. Frequencies of words can be estimated by counting the number of occurrences in the corpus. Each word that occurs in the corpus is counted as an occurrence of each concept containing it.

$$freq(w) = \sum_{w_i \in C_c} count(w_i)$$

where C_c is the set of concepts subsumed by a word *w*. Then, concept probability for *w* can be defined as follows:

$$p(w) = \frac{freq(w)}{N}$$

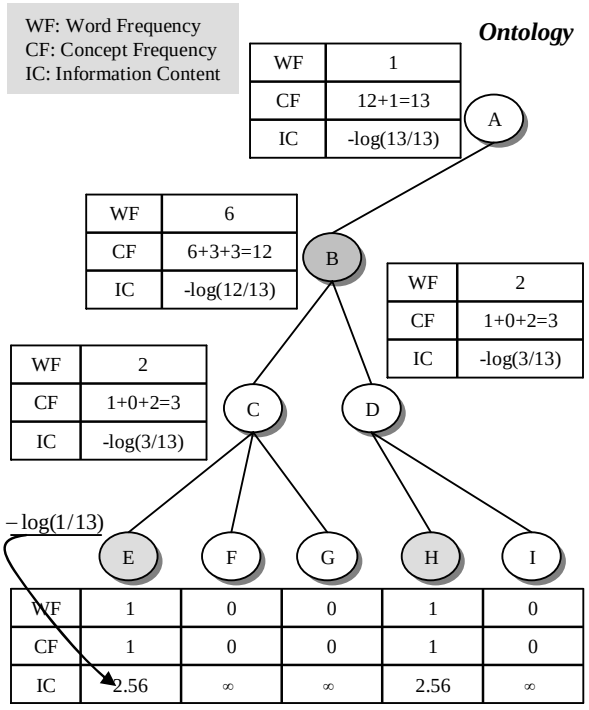


Figure 3. An example of the semantic similarity computation

where N is the total number of words observed in corpus.

This equation states that informativeness decreases as concept probability increases. Thus, the more abstract a concept, the lower its information content. This quantization of information provides a new approach to measure the semantic similarity. The more information that these two words share, the more similar they are. Resnik [7] defines the information that is shared by two words as the maximum information content of the common parents of the words in the ontology.

$$sim_{resnik}(c_i.c_j) = \max_{t \in CP(c_i,c_j)}[-\log p(c)]$$

where $CP(c_i, c_j)$ represents the set of parents words shared by c_i and c_j.

The value of this metric can vary between 0 and infinity. Thus, it is not suitable to use as a probability. Lin [10] suggested a normalized similarity measure as follows:

$$sim\ (c_i.c_j) = \frac{2 \times sim_{resnik}(c_i.c_j)}{-\{\log(p(c_i)) + \log(p(c_j))\}}$$

Figure 3 depicts an instance to compute a similarity between the nodes "E" and "H". The node "B" has the maximum information content of the common parents of the nodes "E" and "H", since the node "B" is the most specific common parent of the nodes "E" and "H". Concept frequency of the node "B" is 12 since it is the sum of its word frequency (6) in the corpus and the sum of the word frequencies (6) of its descendants "C" and "D". Therefore, the similarity between the nodes "E" and "H" is 0.03.

2.2.2. Compound Word Processing

The name of the attributes sometimes consists of a compound word such as "agent name". In English, the meaning of the compound word is generally a spec- ialization of the meaning of its head word. In English, the head word is typically placed on the rightmost position of the word [13, 14]. The modifier limits the meaning of the head word and it is located at the left of the head word [13, 14]. This is most obvious in descriptive compounds, in which the modifier makes it more specific by restricting its scope. A blackboard is a particular kind of board which is black, for instance.

Based on this computational linguistic knowledge, our approach is to give consequence to the mapping with the head word. We discompose the compound word into atomic words and try to compute predicted similarities between each word to the attributes in the other schema. There are two issues of decomposition of the name of the attribute.

- **Tokenization**:
  "agent name" appears in various formats such as "agent_name" or "AgentName". In order to correctly identify these variants, tokenization is applied to names of attributes. Tokenization is a process that identifies the boundaries of words. As a result, non-content bearing tokens (e.g., parentheses, slash, comma, blank, dash, upper case, etc) can be skipped in the matching phase.

- **Stopwords removal**:
  Stopwords are the words that occur frequently in the attribute but do not carry useful information (e.g., of). Such stopwords are eliminated from the vocabulary list considered in the Smart project [15]. Removing the stopwords provides us with flexible matching.

We then integrate each similarity with more weight on the right words.

Let $a_1, a_2, \ldots, a_k$ be a set of tokenized words sorted by rightmost order in the compound word, which is the name of the attribute s in the schema S. The estimated similarity for the s is

$$\sum_{r=1}^{k} \frac{1}{r^2 \cdot N} \times sim(a_r, t), \text{ where } N = \sum_{r=1}^{k} \frac{1}{r^2}$$

r^2 is a heuristic weight value, which is verified in the empirical evaluation. If both attributes are compound words, then let $b_1, b_2, \ldots, b_l$ denote a set of atomic words sorted by the rightmost order in the compound word, which is the name of the attribute t in the schema t. Once we define the similarity between two concepts, the semantic similarity between two attributes (s_i and t_j) can be defined as follows:

$$SIM_{sem}(s_i, t_j) =$$
$$\sum_{q=1}^{l} \frac{1}{q^2 \cdot M} \times \left[\sum_{r=1}^{k} \frac{1}{r^2 \cdot N} \times sim(a_r, b_q) \right]$$
$$where\, M = \sum_{q=1}^{l} \frac{1}{q^2}$$

Thus, the semantic similarity matrix M_{ST}^{sem} is

$$\begin{bmatrix} SIM_{sem}(s_1,t_1) & \cdots & SIM_{sem}(s_1,t_m) \\ \vdots & SIM_{sem}(s_i,t_j) & \vdots \\ SIM_{sem}(s_n,t_1) & \cdots & SIM_{sem}(s_n,t_m) \end{bmatrix}$$

Together with the data-driven mapping frameworks, this framework is optimally combined as described in the next section.

2.3. Similarities Regression

Using the machine learning technique, we combine the predicted similarities: $SIM_{dat}(s_i, t_j)$ and $SIM_{sem}(s_i, t_j)$. Since each similarity can have different significance with contribution to the combined predict-tion, a weight is assigned to each similarity. To impr-ove the predictions of the different single mapping fra-mework, *parameter optimization* [16] is performed where possible by cross-validating on the training data with *logistic regression*. The final estimated similarity between attribute s_i and t_j can be defined as follows:

$$SIM(s_i, t_j) =$$
$$\phi(\alpha \cdot SIM_{dat}(s_i, t_j) + \beta \cdot SIM_{sem}(s_i, t_j))$$
$$where\ \alpha + \beta = 1\ and\ \phi(x) = \frac{1}{1+e^{-x}}$$

Since the *sigmoid* (ϕ) transfers function can divide the whole input space smoothly into a few regions, the desirable prediction can be obtained.

3. Experiments

To demonstrate the accuracy and effectiveness of our mapping framework, we performed experiments on real-world data. We applied our mapping framework to real estate domain datasets that were previously used in LSD [2]. We compare our experiment results with that of complete LSD results in terms of accuracy. The complete LSD matches the schema with the schema and data information.

3.1. Test Datasets

We used two real estate domain datasets. Both datasets contain house for sale listing information. The mediated schema of Real Estate II is larger than that of Real Estate I. **Table 1** illustrates data information of the two datasets.

Domains	Attribute number in the mediated schema	Sources	Downlo-aded listing	Attribut e number in the source schemas	Matchable attribute in the source schemas
Real Estate I	20	5	502-3002	19-21	84-100%
Real Estate II	66	5	502-3002	33-48	100%

Table 1. Domains and data sources for the expreiment [2]

3.2. Experimental Procedure

In order to empirically evaluate our technique, we train the system on the Real Estate I domain (the training domain). With this data, we perform cross-validation ten times to attain more reliable weights for the combination of the predictions from the semantics-driven mapping framework and the data-driven mapping framework. These values are denoted as α and β in Section 2.C.

3.3. Experiment Results

Our experiment aimed to ascertain the relative contributions of utilizing ontologies to identify the semantics of the attributes in the process of schema reconciliation, while LSD exploited learning schema and data information. As shown in **Figure 4**, we have 7.5% and 19.7% higher average accuracy than that of the complete LSD on the two domains.

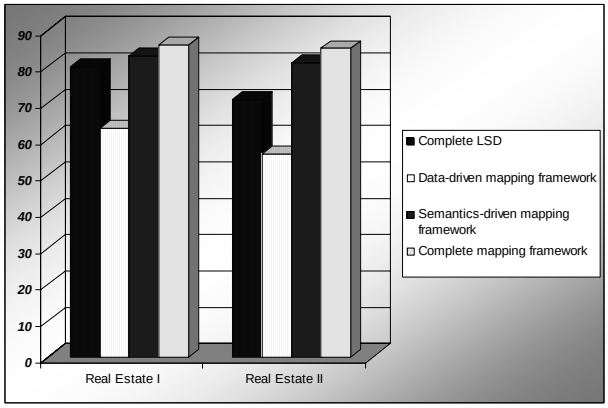


Figure 4. Average matching accuracy comparing with complete LSD [2]

4. Conclusion

We considered the computation of semantic similarity techniques from ontologies to identify the correspondence between database schemas. An experimental prototype system has been developed, implemented, and tested to demonstrate the accuracy of the proposed model which was compared to the previous mapping model. Finally, our future work includes applying this mapping framework into the seismology domain. Seismology data is distributed and organized in different manners and diverse terminologies from various earthquake information providers. This lack of standardization causes problems for seismology research. We anticipate that our framework will successfully resolve this problem.

Acknowledgment

This research was supported by the Computational Technologies Program of NASA's Earth-Sun System Technology Office. The authors first and foremost thank Seokkyung Chung whose contribution in particular has been tremendous. The quality of this work has been vastly improved by his careful and meticulous advice. The authors are also grateful to Sangsu Lee and John O'Donovan.

6. References

[1] R. Dhamankar, Y. Lee, A. Doan, A. Y. Halevy, and P. Domingos, "iMAP: Discovering Complex Mappings between Database Schemas," presented at SIGMOD, 2004.

[2] A. Doan, P. Domingos, and A. Y. Halevy, "Reconciling Schemas of Disparate Data Sources: A Machine-Learning Approach," presented at SIGMOD Conference, 2001.

[3] A. Doan, J. Madhavan, R. Dhamankar, P. Domingos, and Y. Halevy, "Learning to match ontologies on the Semantic Web," *VLDB*, vol. 12, pp. 303-319 2003.

[4] J. Kang and J. Naughton, "On Schema Matching with Opaque Column Names and Data Values," presented at SIGMOD, 2003.

[5] W.-S. Li and C. Clifton, "SEMINT: A tool for identifying attribute correspondence in heterogeneous databases using neural networks," *Data and Knowledge Engineering*, vol. 33, pp. 49-84, 2000.

[6] J. Madhavan, P. Bernstein, A. Doan, and A. Halevy, "Corpus-based Schema Matching," presented at The 21st International Conference on Data Engineering 2005.

[7] 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.

[8] V. I. Levenshtein, "On the Minimal Redundancy of Binary Error-Correcting Codes " *Information and Control* vol. 28, pp. 268-291, 1975.

[9] J. J. Jiang and D. W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," presented at the International Conference on Research in Computational Linguistics, 1998.

[10] D. Lin, "An Information-Theoretic Definition of Similarity," presented at the 15th International Conference on Machine Learning, 1998.

[11] B. Chandrasekaran, J. Josephson, and V. Benjamins, "What are Ontologies, and Why Do We Need Them?," *IEEE Intelligent Systems*, vol. 14, 1999.

[12] Pedersen, Patwardhan, and Michelizzi, "WordNet::Similarity - Measuring the Relatedness of Concepts " presented at the Nineteenth National Conference on Artificial Intelligence (AAAI-04), 2004.

[13] M. Collins, "Three Generative, Lexicalised Models for Statistical Parsing," presented at the 35th Annual Meeting of the ACL (jointly with the 8th Conference of the EACL), 1997.

[14] M. Collins, "A New Statistical Parser Based on Bigram Lexical Dependencies," presented at the 34th Annual Meeting of the ACL, 1996.

[15] G. Salton and M. J. McGill, *Introduction to modern information retrieval*: McGraw-Hill, 1983.

[16] T. Bäck and H.P. Schwefel, *An overview of evolutionary algorithms for parameter optimization*, vol. 1: MIT Press, 1993.