Bridging the gap between data acquisition and inference ontologies – towards ontology based link discovery[*]

Michel L. Goldstein, Steve A. Morris, Gary G. Yen
Oklahoma State University
School of Electrical and Computer Engineering

ABSTRACT

Bridging the gap between low level ontologies used for data acquisition and high level ontologies used for inference is essential to enable the discovery of high-level links between low-level entities. This is of utmost importance in many applications, where the semantic distance between the observable evidence and the target relations is large. Examples of these applications would be detection of terrorist activity, crime analysis, and technology monitoring, among others. Currently this inference gap has been filled by expert knowledge. However, with the increase of the data and system size, it has become too costly to perform such manual inference. This paper proposes a semi-automatic system to bridge the inference gap using network correlation methods, similar to Bayesian Belief Networks, combined with hierarchical clustering, to group and organize data so that experts can observe and build the inference gap ontologies quickly and efficiently, decreasing the cost of this labor-intensive process. A simple application of this method is shown here, where the co-author collaboration structure ontology is inferred from the analysis of a collection of journal publications on the subject of anthrax. This example uncovers a co-author collaboration structures (a well defined ontology) from a scientific publication dataset (also a well defined ontology). Nevertheless, the evidence of author collaboration is poorly defined, requiring the use of evidence from keywords, citations, publication dates, and paper co-authorship.. The proposed system automatically suggests candidate collaboration group patterns for evaluation by experts. Using an intuitive graphic user interface, these experts identify, confirm and refine the proposed ontologies and add them to the ontology database to be used in subsequent processes.

Keywords: ontologies, link discovery, bibliometric analysis, link analysis

1. INTRODUCTION

Most current knowledge discovery applications are focused on the use of massive quantities of data to indirectly infer the occurrence of events that cannot be detected directly. These techniques are now practical because of the decrease in the cost of data storage and processing. Some examples of these growing knowledge inference applications are:

- Detection of terrorism networks (counterterrorism)
- Stock market analysis
- Medical differential diagnosis
- Bibliometric analysis for technology monitoring

In counterterrorism, a highly researched topic in recent years, particularly under the support of DARPA's Information Awareness Office's Evidence Extraction and Link Discovery (EELD) program, the goal is to combine massive, yet sparse data sources, such as telephone records, property ownership, intelligence reports, newspaper articles, scientific reports, which may contain classified and unclassified information, in order to find patterns that suggest illegal activities such as the formation of terrorist groups or the staging of terrorist acts[1]. The stock market has been a very important knowledge inference research area for over a century. The analysis of the prices of the stocks and the volume traded every day can uncover important information about the health of companies and the general state of the market[2]. In the medical differential diagnosis field, the objective is to use diagnosis information, as well as the patient's background and medical history to help in deciding on the disease and treatment[3].

[*] *To appear at the SPIE International Symposium on Aerospace/Defense Sensing, Simulation, and Controls, Orlando, FL, April 21, 2003.*

The study of publications is another important example of the use of large datasets to uncover interesting patterns that do not exist directly in datasets. In this field, by observing published books, scientific papers and other publications, it is possible to detect important researchers, find new research topics, determine important contributors to a field, spot collaboration groups, and investigate interdependencies among research topics[4]. In this paper, journal publications will be used as a testbed for the proposed algorithm because the patterns of interest are well-defined and the also because journal data is well-structured and easily available.

One important commonality between all these applications is that the datasets can be modeled as complex networks with different types of entities and relations between these entities. For example, in the terrorism prevention field, people, places, chemical products, and other physical units can be modeled as entities in a network, and these entities are connected through timed events (meetings, phone calls), infrastructure (location of buildings, automobiles, weapons), and other relations.

We refer to this analysis of data from complex networks to find interesting patterns as *link discovery*. The proposed link discovery process will be explained in more details in the next section. Briefly, it is divided into four main steps:

- **Ontology acquisition:** define the structure of the dataset and the patterns of interest;
- **Link analysis:** identify which relations in the dataset are correlated to the patterns that are being searched for, as most of the relations in the patterns are not directly obtained in the dataset. This creates connections between the patterns of interest and the data structure;
- **Pattern finding:** using the connections defined in the previous step, search for the patterns in the dataset;
- **Results visualization:** display the identified patterns and acquire user feedback regarding the decisions made about the patterns and relations between ontologies.

This paper is centered on the second step of the process, of link analysis. The link analysis process is vital to connect the patterns to the existing relations in the dataset. For example, in the bibliometrics field, the importance of an author is related to the number of that author's publications and, especially, to the number of citations the author's publications receive. The author's institution also is also correlated somewhat to importance. However, there is no logical correlation between keywords used, or date of publication with the high-level property of author importance. The ability to identify properties that are related and unrelated to the given interesting pattern is an important step towards being able to locate the pattern in the dataset. Manual definition of those may be extremely costly in large datasets.

This paper is divided into 6 sections. Section 2 introduces the overall link discovery method. In Section 3 gives a brief introduction to ontology definition and acquisition, an important task that ensures the feasibility of the overall system. Section 4 explains link analysis and proposes a link discovery method. Section 5 gives some experimental results that apply the proposed method finding important relations that indicate author collaboration. Section 6 finalizes with conclusions and directions for future research.

2. LINK DISCOVERY PROCESS

The objective of the link discovery process is to locate and dynamically track previously known patterns, or combination of patterns (scenarios) in a massively large and heterogeneous dataset, and present these patterns to a user for analysis. Note the following definitions:

- **Pattern**: a combination of entities and relations in the data that follow a known constraint. This constraint can be quantitative, qualitative or structural, i.e., a pattern is a combination of entities that share a certain value on one property, or that contain a certain group of properties, or that contains a certain number of a certain type of property, or a combination of these factors.
- **Scenario**: combination of patterns with known structure.

2.1. Proposed link discovery process architecture

A simplified diagram of the link discovery process can be seen below in Figure 1. It is based on four processes, as mentioned earlier, of ontology acquisition, link analysis, link discovery and visualization. The visualization method offers a feedback for all the previous processes, because it may show that the modeling is incorrect, the definition of the important relations is incorrect, or that the identification of potential patterns and the tracking was incomplete.

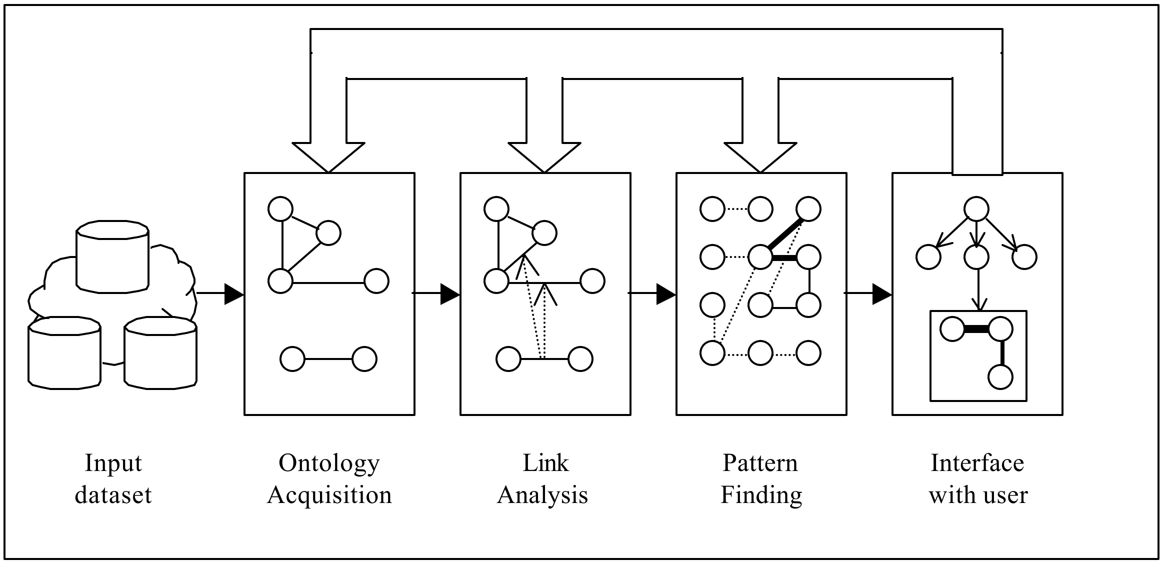

Figure 1 - Simplified Link Discovery Process

The output of this process is a map of the possible entities and relations that form a candidate pattern or scenario. The whole process can be understood with the following simplified example in the medical differential diagnosis domain[3]:

Assume a patient database with the history of various patients regarding diseases, and the background information about those patients. Additionally, assume a comprehensive disease database of known diseases and their relationship to observable symptoms is. The first step is to structure both these datasets in a way that the language used in both are similar and comparable, so that the facts in one database can be mapped to the network in the other database. In the differential diagnosis field, many initiatives exist towards this goal, for example, the Unified Medical Language System (UMLS) by the National Library of Medicine[5] and its subsets, or the Systematized Nomenclature of Medicine (SNOMED) system by the College of American Pathologists[6], or the Clinical Codes used by the British National Health Service[7].

The second step diagnosis is the restructuring of the databases using this common language, or, in a wider sense, a common ontology. Then, when a new patient arrives with an unknown set of symptoms the system can define this patient's ontology, i.e., a network defining the symptoms and background of the patient, as a scenario to look for in the database. The next step of the diagnosis process is to find relations between the ontology of the patient and the ontology of the database. If they were built using the same language, the relation is trivial. However, if a new set of symptoms were observed that could not be unambiguously described in the used ontology, it is necessary to generate indirect connections with the database ontology, i.e. it is necessary to map these observed symptoms to multiple symptoms in the database, possibly including weighting to most similar entities. With these connections made to the structure of the network, i.e. the vocabulary used in the database, a search can be performed by browsing the network of instances of each of the elements. The diagnosis system provides the physician with a map of possible diagnosis and treatments based on similar patterns observed in the disease database.

2.2. Related work

Preliminary research on link discovery is being conducted through DARPA's Evidence Extraction and Link Discovery (EELD) program. The main objective of the EELD initiative is to develop method for identifying possible threat scenarios using classified and unclassified data sources[1]. Researchers at Metron are developing a system that uses Bayesian methods for identifying man-made scenarios[8]. The Experimental Knowledge Systems Laboratory at the University of Massachusetts is applying statistical algorithms to search for relational patterns in temporal data, to cluster data that has similar temporal patterns and, from that, detect interesting scenarios[9]. Holder and Cook from the University of Texas at Arlington are working on a graph-based pattern learning technique based on their well established SUBDUE system for structure discovery[10]. The Naval Center for Applied Research in Artificial Intelligence is working on Analogical Hypothesis Elaborator for Activity Detection (AHEAD), a system for assessing the correctness of hypothesis[11]. It receives preprocessed data and scenario hypothesis and by making analogies and tracing the states of each hypothesis, it calculates the likelihood of it being correct[12].

All these projects are in the initial stages of development. Their main assumption is a well known structure of the input and definition of patterns and scenarios. However, in practical applications, this assumption cannot be safely made. Even experts in a field will have only vague, or high-level abstract mental pictures of the patterns of interest. The connection of these abstract patterns to the actual entities and relations that exist in the database is a very important step towards a more general solution to this wide-reaching problem.

3. ONTOLOGY ACQUISITION

Ontology is the key element to any data processing. It can be defined as "a formal explicit specification of a shared conceptualization."[13] In other words, it is a way to explicitly define the commonalities and differences between the different concepts in a dataset. It defines the semantics of each concept or entity and the way each concept is interrelated, as well as relation constraints.

In the Information Technology field, ontology can be understood as a generic data structure for representing data in a computer-understandable way without loosing human-understandability. Many different languages exist for defining ontologies. Lately, the DAML+OIL (DARPA Agent Markup Language + Ontology Inference Layer) and OWL (Web Ontology Language) are receiving attention from researchers because of their applicability to the Semantic Web. The Semantic Web is a far-reaching long-term project to construct "an extension of the current World Wide Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation."[14] Its objective is to add semantics to web pages and web services in general to make the Web a friendlier place to search and use.

The restructuring of datasets using ontologies, also called ontology generation[15] is based on two main steps: ontology extraction and ontology population. Most researchers consider both processes together, because of their high dependence. However for the growing and heterogeneous datasets associated with complex evolving networks of interest to be investigated using link discovery, it is necessary to focus on ontology extraction and ontology population individually.

3.1. Ontology extraction

Ontology extraction is the process of defining the underlying structure of a dataset. In most of the current ontology applications, this step is done manually using domain experts allied with ontology creation methodologies[16-18] and software[19,20]. These manual methods are very time-consuming, labor-intensive and error-prone, especially when dealing with large domains. Furthermore, ontologies created using manual methods usually lack completeness, i.e., the ontologies obtained using this method are focused on certain elements of the dataset and often oversee relations and elements that look trivial to a human. This lack of completeness can be harmful for the link discovery process because these unaccounted for entities and relations that may be important for discovering the patterns of interest.

There are some important initiatives to automate the process of generating computer-assisted ontologies[21,22]. The most important step to be automated is of defining the hierarchical relations between concepts, i.e. defining possible groups of elements is the dataset that share common properties. These groups form higher-level concepts. The full automation of this process would require the dataset to contain information about how to label and further combine higher-level concepts. Most of datasets do not contain this information, limiting the use of automated methods.

In a journal publications dataset, the ontology is well known and simple allowing manual generation of the ontology as shown in Figure 2.

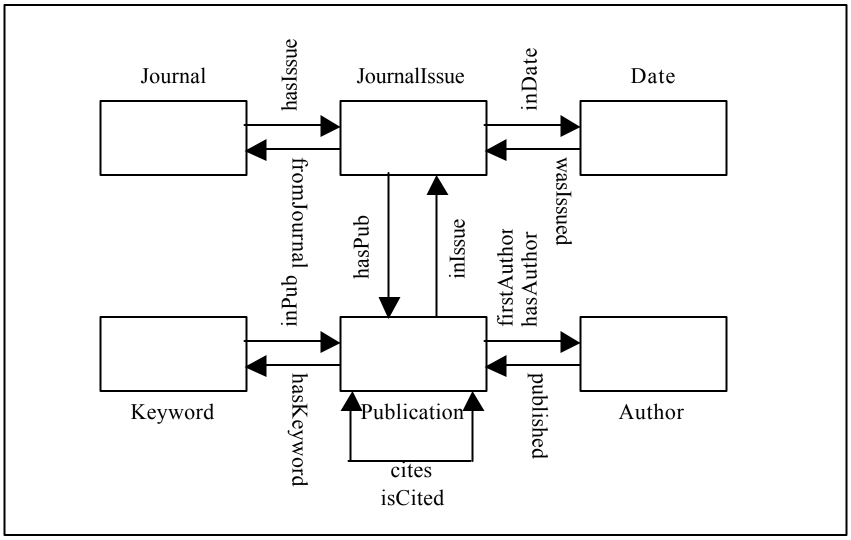

Figure 2 - Simplified view of journal paper collection ontology

3.2. Ontology population

After the ontology is created, the next step is to map the dataset structure to it. This is highly dependent on the way the dataset is obtained. It may include natural language processing (NLP) methods and ontology mapping methods[23].

In the publications dataset to be used as an example in this paper, the source is in a relational database format. Relational database structures can be understood as simple ontologies. Therefore, the population method can be done by mapping this simple ontology to the more comprehensive ontology shown. In order to accomplish the ontology population procedure, it is necessary to relate each field or conjunction of fields in the database to an entity or relation in the ontology. For this simple ontology, it is not a complicated task, however it illustrates the need of general mapping strategies.

4. LINK ANALYSIS

The objective of the link analysis step is to identify relations obtained directly from the database that can be used to infer the strength of some high-level relation. For example, in a publications dataset, if one is looking for author collaboration groups, many different paths in the ontology network can give different connections between authors. Some examples are shown below with some intuition about the meaning of each connection. The examples contain the sequence of relations that connect two authors. For example, authors contain only one outgoing relation (see Figure 2), *published*, therefore all possible paths that connect two authors have to start with this relation.

- published → hasAuthor: direct collaboration by co-authorship
- published → cites → hasAuthor: collaboration by citation of work (weak collaboration)
- published → cites → isCited → hasAuthor: collaboration by co-citation – authors in the same field of work
- published → hasKeyword → inPub → hasAuthor: collaboration by co-use of keywords.

There are many other paths that can be taken. It is important to note that each path taken yields a different interpretation. Thus, any method to identify these indirect relations has to isolate the analysis of each path. Moreover, the coexistence of more than one short paths can also yield important information about the high-level relation.

In order to select the interesting paths or combination of paths, it is necessary to relate them to the pattern of interest. This relationship can occur in three different ways:

- **Explicit characterization when defining the pattern**: when the user defines the pattern of interest, the interesting paths can be defined with it. This is reasonable when the user has good knowledge of the problem. However, when the domain is too large, some relations can be left out because of user oversight;
- **Implicit characterization by pattern examples**: when the knowledge of the pattern is not deep enough, the user can provide examples in the dataset of the patterns of interest and this information can be used to infer the possible interesting paths. However, this solution may be biased because of the chosen examples, and may need, in some applications, a good number of examples to correctly infer the correlation to the interesting paths.
- **Mixed characterization**: generally, a user has partial knowledge of the application and can be able to give general guidelines on the possible interesting paths, or parts of paths. Moreover, obtaining a few examples of patterns may be fairly easy in most domains, and they can be used to reinforce the path suggestions, and complete them.

Having these requirements in mind, a general algorithm will be proposed for defining the interesting paths in determining indirect relationships linked to known interesting patterns.

4.1. Proposed algorithm

Due to the qualitative nature of most datasets, the proposed algorithm was built on qualitative connections only, i.e. only considering binary links. Binary links only contain information of whether two entities are connecter or not. There is no information about weights of connections. Introduction of connections weighting is important when there is uncertainty in the dataset generated by the data acquisition method, such as Natural Language Processing methods, or by the inability for the current ontology to represent the entity and its relations. These weighted links are called quantitative links. Extensions to deal with quantitative links will be discussed in the next subsection, but will not be the focus of the current analysis.

The proposed algorithm is based on an information spread model that assumes that the initial entity under consideration has a known amount of information and that the relations that connect this entity to other entities transmit this information across the network. By explicitly observing the output entity type (from the ontology) from the amount of information transmitted from the initial entity to this entity through chosen relations, or path, it is possible to infer the correlation of the two nodes in the network when used this path. This method is very similar to the Bayesian Belief Network model[24], however, it contains some important differences related to the nature of information transmission that differs from probability transmission in dependency networks. There are similarities also with the work by Girvan and Newman[25], but only one type of relation was considered to generate all results.

There are two distinct types of information transmitting properties: *additive* and *structural*. *Additive* properties are generated when it is assumed that the current information in the node is formed by the combination of the information transmitted from the connecting nodes through that property. For example, an authorship property (*hasAuthor*) is an *additive* property, because usually it can be assumed that the information contained on a publication is a sum of the knowledge from each author. *Structural* properties, on the other hand, are properties that are related to nodes that are connected by the structure of the world, i.e., in an information content sense the two entities are the same, but they are separated in the data structure to ease understanding. Therefore the information content is not divided among the connecting nodes, but only transmitted through them. For example, the other direction of the authorship property, *published*, is a *structural* property, since most of the times it cannot be assumed that authors divide their knowledge into different publications, but simply show it in different ways, especially when dealing with single-subject datasets.

By arbitrarily starting at any particular entity of a given type (based on the pattern of interest), and having a defined set of properties that the information should be transmitted, a correlation can be found to each entity of the target type (also given by the pattern of interest). This forms an information correlation matrix for one set of properties, or path. Applying this method for all sets of paths, a three-dimensional matrix is formed, where for each pair of entities from the source and target types there is a vector of information correlation. This vector can be used to infer the pattern of the connection between these two entities.

From the definition of the pattern that is being searched for, either by examples, explicit definition or by combination, these patterns can be compared and decisions can be taken of whether the found patterns are similar to the target patterns.

The algorithm can be understood in more details in the next subsections, where each of the steps will be explained for each of the different ways of defining a pattern. The experimental results will focus the first two ways of defining a pattern.

4.2. Explicit characterization when defining the pattern

For this example, the definition of the pattern is made by simply defining the paths that should be taken and weight values for each path. The resulting pattern will be a single scalar that is the weighted sum of the distance of each of the given paths:

$$s_{i,j} = \sum_{k=1}^{P} w_k \cdot s_{i,j}^{\;k} \qquad (1)$$

where, $s_{i,j}$ is the similarity between i and j;
w_k is the weight for path k
$s_{i,j}^{k}$ is the similarity between i and j, following path k (see explanation below).

The winner high-level links, i.e., the most probable pair of entities that are interrelated by the pattern of interest, are those paths that that provide the greatest similarity. The choice of working with similarity instead of distance is because two unconnected nodes through one path give an infinite distance. This would bias the whole analysis towards this infinite distance, even when the distance in other paths are low. Since, when performing a weighted sum of the distances between the entities, if one of the distances has infinite distance, i.e., there is no connection between these two entities for the given path, the final distance would be infinite even if there are other paths with very low distance. While using similarity, unconnected nodes receive zero similarity, and that does not change much when calculating the overall similarity (that will be lower because of the zero).

The path similarity $s_{i,j}^{k}$ is computed by getting the total information after the information transmission procedure. The higher the information content on the target node, the higher the similarity, therefore the information is counted as the similarity value.

In networks with qualitative links, one can modify the method for computing the path similarity by weighting each information transmission between entities by the weight of the link used.

4.3. Implicit characterization by providing pattern examples

By providing examples without restricting the paths, it is first necessary to define possible paths. If all relations are considered the same, most of the formed networks are cyclic directed graphs, especially due to the existence of inverse relations, i.e., relations that have a direct opposite meaning to other property. For example, *cites* and *isCitedBy* are inverse relations. Every time one of them is created connecting entities A and B, automatically the other is created connecting entities B and A. Thus, the number of possible paths is infinite. However, the number of meaningful paths is usually finite and they are short. As a result of this, the path defining algorithm uses a simple search method with restricted path length. This method outputs all possible paths that connect two entities in the ontology and has less than a fixed low number (*N*) of properties between them.

After all paths were defined, a similarity vector is produced by calculating the similarity between each pair of entities. Because the distribution of the similarities cannot be considered normal and unimodal, the best way to define other nodes that may display the same similarity is to calculate the Euclidean distance from each similarity vector to each example and use this distance vector (an E-dimensional vector $V_{i,j}$) to make decisions. For this initial analysis, only the norm of this vector is used to generate the decision value. The lower the distance, the better is the choice.

4.4. Mixed characterization

By providing E examples and a set of acceptable paths, or required properties in paths that generates a set of acceptable paths of length N or less, the same process described before is performed. The decrease in the dimensionality of the number of paths decreases the processing time and error caused by the use of meaningless paths.

5. EXPERIMENTAL RESULTS

In this section the results of the explicit scheme for identifying patterns will be shown. The pattern of interest is author collaboration in a single-subject publications dataset. This pattern's ontology is very simple and can be seen in Figure 3, below. As one can observe, the source and target entities are of type Author. The example dataset covers the field of Anthrax research and contains abstracts acquired through the Institute for Scientific Information's Web of Science product have the keywords “Anthrax” or “Anthracis”. A more comprehensive explanation of the example dataset will be shown below.

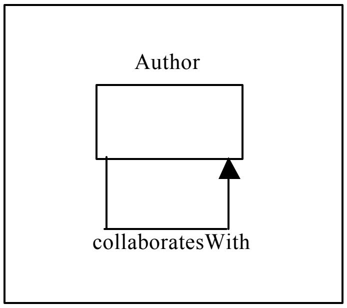

Figure 3 - Author collaboration high-level ontology

5.1. The Anthrax dataset

The Anthrax dataset contains, publications published from January 1945 to the beginning of February 2003, containing the keywords “Anthrax” or “Anthracis”, and their references. This amounts to 2,472 publications and 25,007 unique references, 1,251 of these references correspond to publications in the database. Because by the chosen ontology, references are considered publications, the total number of publications processed is 26,228. As acquired, information on references is limited to their first author, publication year it, the source journal, and the initial page and volume if applicable.

5.2. Author collaboration analysis results by explicit characterization

The first two different methods for defining a pattern were analyzed independently. Results were compared to manual results using timeline maps and crossmaps[26].

Three different explicit characterizations were used independently. They are listed below with their defined path.

- Author collaboration by co-authorship: published → hasAuthor;
- Author collaboration by co-citation: published → cites → isCited → hasAuthor; and
- Author collaboration by use of same keywords: published → hasKeyword → inPub → hasAuthor.

The following tables (Table 1-3) show 10 groups of authors for each of the paths.

Table 1 - Author collaboration by co-authorship results

	Source Author	Target Author
1	Popovic, T	Ashford, DA
2	Popovic, T	Quinn, CP
3	Bhatnagar, R	Ahuja, N
4	Bhatnagar, R	Singh, A
5	Gupta, P	Singh, A
6	Gupta, P	Singh, Y
7	Gupta, P	Bhatnagar, R
8	Gupta, P	Ahuja, N
9	Klein, F	Lincoln, RE
10	Klein, F	Mahlandt, BG

Table 2 - Author collaboration by co-citation results

	Source Author	Target Author
1	Popovic, T	Quinn, CP
2	Bhatnagar, R	Ahuja, N
3	Bhatnagar, R	Singh, Y
4	Bhatnagar, R	Friedlander, AM
5	Gupta, P	Leppla, SH
6	Gupta, P	Bhatnagar, R
7	Gupta, P	Singh, A
8	Gupta, P	Ahuja, N
9	Klein, F	Lincoln, RE
10	Klein, F	Walker, JS

Table 3 - Author collaboration by use of same keywords results

	Source Author	Target Author
1	Bhatnagar, R	Ahuja, N
2	Bhatnagar, R	Singh, Y
3	Bhatnagar, R	Friedlander, AM
4	Gupta, P	Bhatnagar, R
5	Gupta, P	Ahuja, N
6	Mock, M	Fouet, A
7	Mock, M	Montecucco, C
8	Fellows, PF	Little, SF
9	Fellows, PF	Ivins, BE
10	Little, SF	Mock, M

The results obtained by keyword co-use are not as complete as the other methods because only few of the publications in the dataset contain keyword information (263 out of the 26,228). However, this information can be added to the others to generate better identification results.

The good results on fixed paths encourage the investigation of the example-based methods of ontology generation.

5.2. Author collaboration analysis results by implicit characterization using examples

Example author collaboration groups were generated by hierarchical clustering of authors by the number co-authored publications between them. Eighty clusters were generated and five were chosen as examples:

- Quinn, CP and Popovic, T;
- Bhatnagar, R and Gupta, P;
- Bhatnagar, R and Ahuja, N;
- Bhatnagar, R and Singh, A; and
- Klein, F and Lincoln, RE.

Table 4 - Author collaboration by implicit method based on examples

	Source Author	Target Author
1	Mock, M	Fouet, A
2	Mock, M	Schiavo, G
3	Mock, M	Vassaire, J
4	Mock, M	Sirard, JC
5	Mock, M	Guidi-Rontani, C
6	Mock, M	Papini, E
7	Mock, M	Mesnage, S
8	Mock, M	Labruyere, E
9	Mock, M	Munier, H
10	Mock, M	Lereclus, D

By automatically generating all ten possible paths between two authors with distance 4 or less (greater distances lose meaning), one can apply the proposed pattern recognition method to identify author collaborations. Some collaboration groups obtained can be seen in Table 4.

Most of these collaboration pairs are confirmed by the clustering method.

6. CONCLUSIONS

The proposed algorithm has shown to be very powerful for establishing the correlation between elements that can indicate indirect relationships between these elements. Moreover, by using examples, one can use these correlations through different paths in the network as features to indicate desired indirect relationships. The ability to combine the distances through different paths to generate desired features is one of the most important differences of this framework to previous methods.

The use of ontologies to represent the structure of the dataset and the structure of the patterns being searched for has added a powerful human-understandability factor to the method. This renders feasible the use of composed features in complex networks, because it helps the identification of semantically incorrect paths for establishing information correlation. Moreover, ontologies offer a general method for structuring datasets enabling the creation of more general methods for inferring indirect relationships.

One of the greatest challenges when dealing with large datasets is of visualizing the results so that it is possible to interpret and act upon the results quickly and efficiently. The hierarchical nature of ontologies is very helpful to enable a more natural organization of the results. However, the datasets that are freely available do not contain this hierarchical information. There is a tendency for this information to be added to all databases, but, due to the large legacy systems that exist, the timeframe for this change becomes prohibitive for the development of this essential technology. Therefore, hierarchy-generating methods, such as hierarchical clustering, has to be applied on the structure itself to enable visualization.

Finally, link analysis is a small but essential stage of the whole link discovery process. When the whole process is defined and implemented, it has to deal with arbitrarily large datasets. The current method of information transmission does not offer good scalability properties, because it requires the storage of the whole ontology in memory to enable the information transmission process. This same problem has been seen in other network-based methods, such as Bayesian Belief Networks. Currently there is ongoing research on finding scalable methods that can approximate the result of these network spread methods. Markov Chain Monte Carlo methods[27] are being highly researched lately as a potential method to deal with this issue.

REFERENCES

1. T. Senator, "Evidence extraction and link discovery," DARPA Information Awareness Office, http://www.darpa.mil/iao, 2001.
2. S. Siekmann, R. Kruse, J. Gebhardt, F. van Overbeek and R. Cooke, "Information fusion in the context of stock index prediction," *International Journal of Intelligent Systems*, **16**(11), pp. 1285-1298, 2001.
3. E.H. Shortliffe, "Clinical decision-support systems," *Medical Informatics: Computer Applications in Health Care*, E.H. Shortliffe, L.E. Perreult, G. Wiederhold and L.M. Fagan, editors, pp. 466-502, Addison-Wesley, Reading, MA, 1990.
4. H. D. White and K. W. McCain, "Bibliometrics," *Annual Review of Information Science and Technology*, **24**, pp. 119-186, 1989.
5. B. L. Humphreys, D. A. B. Lindberg, H. M. Schoolman and G. O. Barnett, "The unified medical language system: and informatics research collaboration," *Journal of the American Medical Informatics Association*, **5**(1), pp. 1-11, 1998.
6. College of American Pathologists, "SNOMED: systematized nomenclature of medicine," http://www.snomed.org, 2001
7. NHS Information Authority, "Clinical terminology services," http://www.nhsia.nhs.uk/terms/pages/default.asp, 2002.
8. Metron Inc., "Evidence extraction and link discovery program," http://vivaldi.metsci.com/eeld/index.html, 2002.
9. P. R. Cohen, T. Oates, N. Adams and C. R. Beal, "Robot baby 2001," *Proceedings of the 12th International Conference on Algorithmic Learning Theory*, pp. 32-56, Washington, DC, 2001.
10. I. Jonyer, L. B. Holder, D. J. Cook, "Hierarchical conceptual structural clustering," *International Journal on Artificial Intelligence Tools*, **10**(1-2), pp. 107-136, 2001.
11. D. W. Aha, L. A. Breslow and H. Muñoz-Avila, "Conversational base-based reasoning," *Applied Intelligence*, **14**(1), pp. 9-32, 2001.
12. J. W. Murdock, D. W. Aha and L. A. Breslow, "AHEAD: case-based process model explanation of asymmetric threats," *Technical Report AIC-02-203*, Naval Research Laboratory, Navy Center for Applied Research in Artificial Intelligence, Washington, DC, 2002.
13. T. R. Gruber, "A translation approach to portable ontology specifications," *Knowledge Acquisition*, **5**(2), pp. 199-220, 1993.
14. T. Berners-Lee, J. Handler and O. Lassila, "The Semantic Web - A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities," *Scientific American*, **284**(5), pp. 34-43, 2001.
15. Y. Ding and S. Foo, "Ontology research and development. Part 1 – A review of ontology generation," *Journal of Information Science,* **28**(2), pp. 123-136, 2002.
16. M. Uschold and M. Gruninger, "Ontologies: principles, methods and applications," *Knowledge Engineering Review*, **11**(2), pp. 93-136, 1996.
17. M. Fernández López, "Overview of methodologies for building ontologies," *Proceedings for IJCAI-Workshop on Ontologies and Problem-Solving Methods*, pp. 26-34, Stockholm, Sweden, 1999.
18. N. Guarino, "Some ontological principles for designing upper level lexical resources," *Proceedings ofr the 1st International Conference on Lexical Resources and Evaluation*, pp. 527-534, Granada, Spain, 1998.
19. S. Bechhofer, I. Harrocks, C. Goble and R. Stevens, "OilEd: a reasonable ontology editor for the semantic web," *Proceedings of the Joint German/Austrian Conference on Artificial Intelligence*, pp. 369-408, Vienna, Austria, 2001.
20. N. F. Noy, M. Sintek, S. Decker, M. Crubézy, R. W. Fergerson and M. A. Musen, "Creating semantic web contents with Protégé-2000," *IEEE Intelligent Systems*, **16**(2), pp. 60-71, 2001.

21. J. Jannik and G. Wiederhold, “Ontology maintenance with an algebraic methodology: a case study,” *Proceedings of the AAAI Workshop on Ontology Management*, pp. 40-47, Orlando, FL, 1999.
22. A. Maedche and S. Staab, “Discovering conceptual relations from text,” *Proceedings of the 14th European Conference on Artificial Intelligence*, pp. 20-25, Amsterdam, The Netherlands, 2000.
23. Y. Ding and S. Foo, “Ontology research and development. Part 2 – A review of ontology mapping and evolving,” *Journal of Information Science*, **28**(5), pp. 375-388, 2002.
24. N. Friedman and D. Koller, “Being Bayesian about network structure: a Bayesian approach to structure discovery in Bayesian networks,” *Machine Learning*, **50**(1-2), pp. 95-125, 2003.
25. M. Girvan, M. E. J. Newman, “Community structure in social and biological networks,” *Proceedings of the National Academy of Sciences of the United States of America*, **99**(12), pp. 7821-7826, 2002.
26. S. A. Morris, G. G. Yen, Z. Wu and B. Asnake, “Timeline visualization of research fronts,” *Journal of the American Society for Information Science and Technology*, **54**(5), pp. 413-422, 2003.
27. C. Andrieu, N. Freitas, A. Doucet and M. I. Jordan, “An Introdution to MCMC for Machine Learning,” *Machine Learning*, **50**(1-2), pp. 5-43, 2003.