International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue V, May 2019- Available at www.ijraset.com

Survey of Hyponym Relation Extraction from Hyperlinks using Motif Patterns with Feature Combination Extraction Model

G. Keerthiga[1], P. Mallika[2], A. Gokilavani[3]
[1] *ME 4th Semester/ Department of CSE/Jai Shriram Engineering College/Tirupur/TN*
[2, 3] *Assistant Professor/prof/Department of CSE/Jai Shriram Engineering College/Tirupur/TN*

Abstract: ***This paper presents a method for measuring the semantic similarity between concepts in Knowledge Graphs (KGs) such as WordNet and DBpedia. Previous work on semantic similarity methods have focused on either the structure of the semantic network between concepts (e.g., path length and depth), or only on the Information Content (IC) of concepts. We propose a semantic similarity method, namely wpath, to combine these two approaches, using IC to weight the shortest path length between concepts. Conventional corpus-based IC is computed from the distributions of concepts over textual corpus, which is required to prepare a domain corpus containing annotated concepts and has high computational cost. As instances are already extracted from textual corpus and annotated by concepts in KGs, graph-based IC is proposed to compute IC based on the distributions of concepts over instances. Measuring the similarity between documents is an important operation in the text processing field. This project proposed a new similarity measure. Discovering hyponym relations among domain-specific terms is a fundamental task in taxonomy learning and knowledge acquisition. However, the great diversity of various domain corpora and the lack of labeled training sets make this task very challenging for conventional methods that are based on text content. The hyperlink structure of Wikipedia article pages was found to contain recurring network motifs in this study, indicating the probability of a hyperlink being a hyponym hyperlink. Hence, a novel hyponym relation extraction approach based on the network motifs of Wikipedia hyperlinks was proposed. This approach automatically constructs motif-based features from the hyperlink structure of a domain; every hyperlink is mapped to a 13-dimensional feature vector based on the 13 types of three-node motifs.***

Keywords: Motif Relationship, Hyponym Relationship, Semantic Similarity,

I. INTRODUCTION

Data mining is the process of extracting patterns from database. Data mining is seen as increasingly important tool by modern business to transform data into an informational advantage. It is used in a profiling practices, such as marketing, surveillance, fraud detection, and scientific discovery. These techniques can however, be used in the creation of new hypothesis to test against the larger data populations

Data mining derives its name by finding the similarities between searching for valuable information in a large database.. Document clustering is being studied from many decades but still it is far from a trivial and solved problem.

A. Identifying sentences that provide the information regarding success factors and their relationships by utilizing data mining technique
B. The optimal workflow derived from the results by similar classification task
C. Selecting appropriate features of the documents that should be used for clustering.
D. Selecting an appropriate similarity measure between documents.
E. Implementing the clustering algorithm that makes it feasible in terms of required memory and CPU resources.
F. Finding ways of assessing the quality of the clustering technique.

Discovering hyponym relations among domain-specific terms is a fundamental task in taxonomy learning and knowledge acquisition. However, the great diversity of various domain corpora and the lack of labeled training sets make this task very challenging for conventional methods that are based on text content. The hyperlink structure of Wikipedia article pages was found

to contain recurring network motifs in this study, indicating the probability of a hyperlink being a hyponym hyperlink. Hence, a novel hyponym relation extraction approach based on the network motifs of Wikipedia hyperlinks was proposed. This approach automatically constructs motif-based features from the hyperlink structure of a domain; every hyperlink is mapped to a 13-dimensional feature vector based on the 13 types of three-node motifs. The approach extracts structural information from Wikipedia and heuristically creates a labeled training set. Classification models were determined from the training sets for hyponym relation extraction. Two experiments were conducted to validate our approach based on seven domain-specific datasets obtained from Wikipedia.

The first experiment, which utilized manually labeled data, verified the effectiveness of the motif-based features. The second experiment, which utilized an automatically labeled training set of different domains, showed that the proposed approach performs better than the approach based on lexico-syntactic patterns and achieves comparable result to the approach based on textual features. The term hyponym indicates "a-type-of" relationship. For example, maple is a hyponym of tree,and tree is a hyponym of plant. If lexical term ti is the hyponym of another lexical term tj, then ti and tj have a hyponym relation. The hyponym relation is a fundamental type of semantic relation connecting a wide variety of concepts or domain-specific terms to form a semantic taxonomy. Hyponym relation extraction from Web pages plays a crucial role in web information extraction, taxonomy learning, knowledge acquisition, and other knowledge-rich problems.

Wikipedia has become a popular data source in hyponym relation extraction research. Several such studies adopted the syntactic-pattern-based methods or textural- feature-based machine learning methods to extract hyponym relations from Wikipedia. These methods rely mainly on features extracted from the text content of Wikipedia. When shifting to a new domain, these methods require new syntactic patterns to be learned or new training samples to be manually constructed, which usually entail high labor costs. In addition, these methods do not fully utilize the topological structure of hyperlinks in Wikipedia article pages. This paper considers the topological structure of Wikipedia hyperlinks as an important type of feature in hyponym relation extraction.

II. RELATED WORKS

U. Alon [1] describe the main idea that is presented in this Review is that each network motif can carry out specific information-processing functions. These functions have been analysed using mathematical models and tested with dynamic experiments in living cells. Still, there is much to be done: it is important to further experimentally test the functions that each network motif can perform. Such experiments could illuminate the dynamics of the many systems in which each motif appears. Furthermore, it is important to test whether motifs can help us to understand the densely connected networks of higher organisms.

C. Andrew et al [2] considered here the problem of building a never-ending language learner; that is, an intelligent computer agent that runs forever and that each day must (1) extract, or read, information from the web to populate a growing structured knowledge base, and (2) learn to perform this task better than on the previous day. In particular, we propose an approach and a set of design principles for such an agent, describe a partial implementation of such a system that has already learned to extract a knowledge base containing over 242,000 beliefs with an estimated precision of 74% after running for 67 days, and discuss lessons learned from this preliminary attempt to build a never-ending learning agent. The paper underlying this research is that the vast redundancy of information on the web (e.g., many facts are stated multiple times in different ways) will enable a system with the right learning mechanisms to succeed. One view of this research is that it is a case study in lifelong, or never-ending learning. A second view is that it is an attempt to advance the state of the art of natural language processing. A third view is that it is an attempt to develop the world's largest structured knowledge base – one that reflects the factual content of the world wide web, and that would be useful to many AI efforts.

N. Y. Asuka Sumida [3] describes an extension of Sumida and Torisawa's method of acquiring hyponymy relations from hierachical layouts in Wikipedia (Sumida and Torisawa, 2008). We extract hyponymy relation candidates (HRCs) from the hierachical layouts in Wikipedia by regarding all subordinate items of an item x in the hierachical layouts as x's hyponym candidates, while Sumida and Torisawa (2008) extracted only direct subordinate items of an item x as x's hyponym candidates. We then select plausible hyponymy relations from the acquired HRCs by running a filter based on machine learning with novel features, which even improve the precision of the resulting hyponymy relations. Experimental results show that we acquired more than 1.34 million hyponymy relations with a precision of 90.1%. The goal of this study is to automatically extract a large set of hyponymy relations, which play a critical role in many NLP applications such as Q&A systems (Fleischman et al., 2003) and specification retrieval (Yoshinaga and Torisawa, 2006). In this paper, a hyponymy relation is defined as a relation between a hypernym and a hyponym when "the hyponym is a (kind of) hypernym." We acquired more than 1.34 million hyponymy relations in Japanese with a precision of 90.1%.

Sergey Chernov [4] propose two measures for automatic filtering of strong semantic connections between Wikipedia categories. One measure is the number of links between pages in two categories, and the other is Connectivity Ratio. They can be applied to inlinks or outlinks separately. For evaluation, we apply these measures to the English Wikipedia and perform user study to assess how semantically strong the extracted relationships are. We observe that both number of links and Connectivity Ratio correlates with semantic connection strength. It supports our hypothesis, while much more experiments are needed to achieve a convincing evaluation.

Michele Banko [5] explained about Information Extraction (IE) has focused on satisfying precise, narrow, pre-specified requests from small homogeneous corpora (e.g., extract the location and time of seminars from a set of announcements). Shifting to a new domain requires the user to name the target relations and to manually create new extraction rules or hand-tag new training examples. This manual labor scales linearly with the number of target relations. This paper introduces Open IE (OIE), a new extraction paradigm where the system makes a single data-driven pass over its corpus and extracts a large set of relational tuples without requiring any human input. The paper also introduces TEXTRUNNER, a fully implemented, highly scalable OIE system where the tuples are assigned a probability and indexed to support efficient extraction and exploration via user queries.

Johannes Hoffart Fabian [6] presents an automatic approach to the construction of BabelNet, a very large, wide coverage multilingual semantic network. Key to our approach is the integration of lexicographic and encyclopedic knowledge from WordNet and Wikipedia. In addition, Machine Translation is applied to enrich the resource with lexical information for all languages. We first conduct in vitro experiments on new and existing gold-standard datasets to show the high quality and coverage of BabelNet. We then show that our lexical resource can be used successfully to perform both monolingual and cross-lingual Word Sense Disambiguation: thanks to its wide lexical coverage and novel semantic relations, we are able to achieve state-of the-art results on three different SemEval evaluation tasks.

III. METHODOLOGY

A. Wag Construction

This section constructs the Wikipedia Article Graph. During this, the pages referred to external web pages or files not found in this collection are eliminated. Using regular expression, the pages are found out using the links. Then a graph is formed such that the pages are being nodes and the links are being edges. This considers the topological structure of Wikipedia hyperlinks as an important type of feature in hyponym relation extraction. Each Wikipedia article page represents a domain-specific term.

It contains a number of hyperlinks pointing to other article pages. Fig. 3.1 shows a fragment of article page k-medoids algorithm, which contains three hyperlinks. These hyperlinks and article pages can be considered as directed graphs. Then these graphs are named as the Wikipedia article graphs (WAGs). The hyperlinks in a WAG can imply semantic relations such as hyponymy relation between the two connected article pages.

.

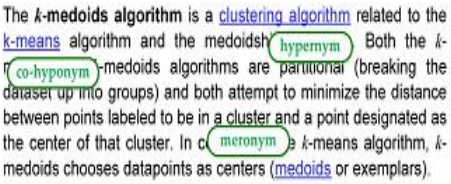

Fig.3.1 Fragment of a Wikipedia article page that illustrates the semantics of Wikipedia hyperlinks.

For example, in Fig. 1 the first hyperlink indicates that K-mediods algorithm is a hyponym of clustering algorithm, and the second hyperlink indicates that K-mediods algorithm is a co-hyponymy of k-means.

$$HHR(j) = \frac{\#\text{ hyponym hyperlinks contained by the instances of motif } j}{\#\text{ all hyperlinks contained by the instances of motif } j}. \quad (2)$$

B. Data Cleaning

1) *Stem Word:* In this module, the word and its stem word is keyed in and saved into the table. The details are saved in 'Stemword' table.
2) *Stop Word:* In this module, the stop word is keyed in and saved into the table. The detail is saved in 'Stopword' table.

3) *Synonym Word:* In this module, the word and its synonym word is keyed in and saved into the table. The details are saved in 'Synonym' table.
4) *Hypnonym Word:* In this module, the word and its hyponym word is keyed in and saved into the table. The details are saved in 'Hyponym' table.
5) *Preprocessing:* In this module, all the documents downloaded are applied with stemming, stop word removal and synonym word replacement.

C. Motif Pattern Construction

In this section, the three node network motif patterns are constructed. The two parameters below were utilized to qualify the three-node motifs.

1) Z-Score indicates the statistical significance of a network motif. The Z-Score of motif j is formally defined in (1).

$$Z\text{-}Score(j) = \frac{N(j) - \overline{N_r(j)}}{\sigma_r(j)}, \qquad (1)$$

where N(j) is the number of occurrences of motif j($1 \leq j \leq 13$) in network N.Nr(j) is the average number of occurrences of motif j in an ensemble of randomized networks with the same degree of distribution as network N. σr(j) is the standard deviation of Nr(j). In general, a motif with a high Z-Score indicates that the motif appears in a particular network (N) more frequently than in randomized networks.

2) A new parameter, Hyponym Hyperlink Rate (HHR), was introduced to describe the sparsity of hyponym relations within a network motif. The HHR of motif j is defined in (2). The higher the HHR of a network motif is, the denser the hyponym hyperlinks in the motif are. This condition means that if a hyperlink appears in a motif with high HHR, then this hyperlink is likely to be a hyponym hyperlink.

All the links which are related with hyponym are considered in our project.

D. Feature Based Text Content

In this module, the features based on text content are also combined. For example, the word mobile if contained, it relates the links of pages containing the <company name> mobile and other mobile related phrases even if does not behave as hyponym for same context.

In proposed system, the study presents the construction of the domain-specific datasets from the Wikipedia hyperlinks as follows

1) For each of the seven domains, the existing system crawls the Wikipedia article pages from the start position to a depth of 3. With the domain Data mining as an example, it crawls the article pages by traversing article-article hyperlinks from the Data mining article page.
2) A set of URL regular expressions was utilized during crawling to remove irrelevant article pages, such as External links and Languages.
3) The three-node motifs in these datasets were analyzed based on the results which can quickly identify network motifs from large graphs with the data structure.

The column "#Instances" indicates the total number of three-node motif instances. The two parameters below were utilized to qualify the three-node motifs.

$$Z\text{-}Score(j) = \frac{N(j) - \overline{N_r(j)}}{\sigma_r(j)}, \qquad (1)$$

a) Z-Score indicates the statistical significance of a network motif. The Z-Score of motif j is formally defined in (1). where N(j) is the number of occurrences of motif j($1 \leq j \leq 13$) in network N.Nr(j) is the average number of occurrences of motif j in an ensemble of randomized networks with the same degree of distribution as network N. σr(j) is the standard deviation of Nr(j). In general, a motif with a high Z-Score indicates that the motif appears in a particular network (N) more frequently than in randomized networks.
b) A new parameter, Hyponym Hyperlink Rate (HHR), was introduced to describe the sparsity of hyponym relations within a network motif. The HHR of motif j is defined in (2). The higher the HHR of a network motif is, the denser the hyponym hyperlinks in the motif are. This condition means that if a hyperlink appears in a motif with high HHR, then this hyperlink is likely to be a hyponym hyperlink.

$$HHR(j) = \frac{\#\text{ hyponym hyperlinks contained by the instances of motif } j}{\#\text{ all hyperlinks contained by the instances of motif } j}. \qquad (2)$$

E. Advantages

In addition with all the existing system mechanism, the proposed study includes stemming, stop word removal and synonym word replacement. The features based on text content are also combined. For example, the word mobile if contained, it relates the links of pages containing the <company name> mobile. Most occurring domain-specific terms containing more different words are also taken for label setting.

1) Data preprocessing steps as Stemming, stop words removal and synonym word replacement is also considered.
2) This approach may work well in a domain where the hyponym relations among domain-specific terms are containing more different words for same meaning.
3) The features based on text content are also combined. For example, the word mobile if contained, it relates the links of pages containing the <company name> mobile and other mobile related phrases even if does not behave as hyponym for same context

F. Automatic Training Set Algorithm (Atsa)

1) LET D_{List} = Wiki Documents List Downloaded
2) PositiveSet = {}
3) NegativeSet = {}
4) For each Document d in DList
5) S =content(d)
6) $D_{Others = DList \setminus d}$
7) Fetch outgoing links E if S contains link for documents in D_{Others}
8) If sizeof(E)= 1 Then
9) I = Find Incoming edges of ‘d’
10) If I=1 Then
11) NegativeSet = NegativeSet □□d
12) Else
13) If Not d □□NegativeSet Then
14) PositiveSet = PositiveSet □□d
15) End If
16) End If
17) End If

G. Wiki Article Graph Construction

1) First wiki documents are downloaded from the internet and saved in the name of the hyperlink they point to.
2) The document file details are saved as nodes in ‘URL’ table. From which positive set documents are found out using ATSA Algorithm. These are referred as **D_{List}**
3) For each Document d in DList
4) S= content(‘d’)
5) $D_{Others = DList \setminus ‘d’}$
6) Fetch links E if S contains link for documents in D_{Others}
7) Add the links E to d. These become outgoing edges of ‘d’.
8) Next

H. Motif Pattern Construction

1) For each Document d in DList
2) S= content(‘d’)
3) $D_{Others = DList \setminus ‘d’}$
4) Fetch Incoming links with hynonym word $E_{Incoming}$ if S contains link for documents in D_{Others}
5) Fetch Outgoing links with hynonym word $E_{outgoing}$ if S contains link for documents in D_{Others}
6) Find all motif patterns M such that each motif pattern contains d and any other two nodes fall in either $E_{Incoming}$ and e $_{ing}$ if Z-Score value is more so that it is considered as a significant motif.

I. Motif Pattern Construction

1) For each Document d in DList

2) S= content('d')

3) D_{Others} = DList \ 'd'

4) Fetch Incoming links with synonym word (Even if it not hyponym) $E_{Incoming}$ if S contains link for documents in D_{Others}

5) Fetch Outgoing links with synonym word (Even if it not hyponym) $E_{outgoing}$ if S contains link for documents in D_{Others}

6) Find all motif patterns M such that each motif pattern contains d and any other two nodes fall in either $E_{Incoming}$ and $E_{outgoing}$ if Z-Score value is more so that it is considered as a significant motif.

IV. CONCLUSION

This paper focuses on extracting hyponym relations from academic domains like Wikipedia web sites. Such Wikipedia data contain a certain amount of hyponym relations between domain-specific terms. It is showed that the motif-based features are effective in hyponym relation extraction. The experiment results showed that the proposed system approach works well in domains where there are enough hyponym relations. Since the existing approach may not work well in a domain where the hyponym relations among domain-specific terms are very sparse, such as human individuals or companies, a feature based motif pattern construction is also added to solve the problem.It also proposes features based on text content combined with network motifs to improve extraction performance. The experiments are conducted after features are combined with normal hyponym words and are tested. The application can be used to extract motif patterns from any knowledge based articles web site.

REFERENCES

[1] B. Nath, D. Bhattacharyya, and A. Ghosh, "Discovering association rules from incremental datasets," IJCSC, vol. 1, no. 2, pp. 433–441, 2010
[2] Y. Cao, H. He, and H. Man, "SOMKE: Kernel density estimation over data streams by sequences of self-organizing maps," IEEE Trans. Neural Netw. Learn. Syst., vol. 23, no. 8, pp. 1254–1268, Aug. 2012.
[3] P. Domingos and G. Hulten, "A general framework for mining massive data stream," J. Comput. Graphical Statist., vol. 12, no. 4, pp. 945–949, 2003
[4] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, "A framework for on-demand classification of evolving data streams," IEEE Trans. Knowl. Data Eng., vol. 18, no. 5, pp. 577–589, May 2006
[5] B. Babcock, S. Babu, M. Datar, R. Motwani, and J. Widom, "Models and issues in data stream systems," in Proc. 21st ACM Symp. Principles Database Syst., 2002, pp. 1–16
[6] B. W. Silverman, Density Estimation for Statistics and Data Analysis. London, U.K.: Chapman & Hall, 1986
[7] J. DiNardo and J. L. Tobias, "Nonparametric density and regression estimation," J. Economic Perspectives, vol. 15, no. 4, pp. 11–28, 2001
[8] Y. Cao, H. He, H. Man, and X. Shen, "Integration of self-organizing map (som) and kernel density estimation (kde) for network intrusion detection," Proc. SPIE, vol. 7480, pp. 74800N-1–74800N-12, Sep. 2009
[9] T. Brox, B. Rosenhahn, D. Cremers, and H. P. Seidel, "Nonparametric density estimation with adaptive, anisotropic kernels for human motion tracking," in Proc. 2nd Conf. Human Motion Underst. Model. Capture Animation, 2007, pp. 152–165
[10] T. Bouezmarni and J. V. K. Rombouts, "Nonparametric density estimation for multivariate bounded data," J. Statist. Plann. Inference, vol. 140, no. 1, pp. 139–152, 2010
[11] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: A collaboratively created graph database for structuring human knowledge," in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2008, pp. 1247–1250.
[12] C. Bizer, et al., "Dbpedia-a crystallization point for the web of data," Web Semantics: Sci. Services Agents World Wide Web, vol. 7, no. 3, pp. 154–165, 2009.
[13] J. Hoffart, F. M. Suchanek, K. Berberich, and G. Weikum, "Yago2: A spatially and temporally enhanced knowledge base from wikipedia (extended abstract)," in Proc. 23rd Int. Joint Conf. Artif. Intell., 2013, pp. 3161–3165.
[14] R. Navigli and S. P. Ponzetto, "Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network," Artif. Intell., vol. 193, pp. 217–250, 2012.
[15] E. Hovy, R. Navigli, and S. P. Ponzetto, "Collaboratively built semi-structured content and artificial intelligence: The story so far," Artif. Intell., vol. 194, pp. 2–27, 2013.
[16] A. Moro, A. Raganato, and R. Navigli, "Entity linking meets word sense disambiguation: A unified approach," Trans. Assoc. Comput. Linguistics, vol. 2, pp. 231–244, 2014.
[17] J. Hoffart, S. Seufert, D. B. Nguyen, M. Theobald, and G. Weikum, "Kore: Keyphrase overlap relatedness for entity disambiguation," in Proc. 21st ACM Int. Conf. Inf. Knowl. Manage., 2012, pp. 545–554.
[18] I. Hulpus, N. Prangnawarat, and C. Hayes, "Path-based semantic relatedness on linked data and its use to word and entity disambiguation," in Proc. 14th Int. Semantic Web Conf., 2015, pp. 442–457.
[19] M. Schuhmacher and S. P. Ponzetto, "Knowledge-based graph document modeling," in Proc. 7th ACM Int. Conf. Web Search Data Mining, 2014, pp. 543–552.
[20] S. Shekarpour, E. Marx, A.-C. N. Ngomo, and S. Auer, "Sina: Semantic interpretation of user queries for question answering on interlinked data," Web Semantics: Sci. Services Agents World Wide Web, vol. 30, pp. 39–51, 2015.