International Journal of Recent Trends in Engineering & Research
ISSN : 2455-1457 Impact Factor : 3.344

WORD SENSE DISAMBIGUATION BASED ON SUPERVISED MODEL

Nithya. S.P[1], Nithya. S[2], Malleshkumar. M[3]
[1,2,3]Student, Department of CSE, K.S.Rangasamy College of Technology, Tiruchengode, Tamil Nadu, India

Abstract: Measuring the semantic similarity between words is an important component in various tasks on the web are relation extraction, community mining, document clustering, and automatic metadata extraction.The usefulness of semantic similarity measures in these applications, accurately measuring semantic similarity between two words remains a challenging task. The project proposes an empirical method to estimate semantic similarity using page counts and text snippets retrieved from a web search engine for two words. Specifically, it defines various word co-occurrence measures using page counts and integrates those with lexical patterns extracted from text snippets. To identify the various semantic relations that exist between two given words, the project proposes a novel pattern extraction.

I. INTRODUCTION

Data mining an interdisciplinary subfield of computer science, is the computational process of discovering patterns in a large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall aim of the data mining process is to extract information from a data set and transform it into an understandable structure for further usage. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and the inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization , and in online updating

Data mining involves six common classes of tasks: Anomaly detection the identification of unusual data records, that is interesting or data errors that require further investigation.

Association rule learning (Dependency modeling) – Searches for relationship between variables. For example a supermarket might gather data on customer purchasing habits. Using association rules learning, the supermarket can determine which products are frequently bought together and use this information for marketing purpose. This is sometimes referred to as market basket analysis.

Clustering – is the task of discovering groups and structure in the data that are in some way or another "similar", without using known structures in the data.

Classification – the task of generalizing known structure to apply to new data. For example, an e-mail program would attempt to classify an e-mail as "legitimate" or as "spam".

Regression – attempts to find a function which models the data with the least error.

Summarization – providing a more compact representation of the data set, including visualization and report generation.

II. EXISTING SYSTEM

Accurately measuring the semantic similarity between words is an important problem in web mining information retrieval, and natural language processing. Web mining applications such as, community extraction relation detection, and entity disambiguation, require the ability to accurately measure the semantic similarity between concepts or entities. In information retrieval, one of the main problems is to retrieve a set of documents that is semantically related to a given user query. Efficient estimation of

semantic similarity between words is critical for various natural language processing tasks such as word sense disambiguation (WSD) and automatic text summarization.

Semantically related words of a particular word are listed in manually created general-purpose lexical ontologies such as WordNet.In WordNet, a synset contains a set of synonymous words for a particular sense of a word. However, semantic similarity between entities changes over time and across domains. For example, apple is frequently associated with computers on the web. However, this sense of apple is not listed in most general-purpose thesauri or dictionaries.A user, who searches for apple on the web, might be interested in this sense of apple and not apple as a fruit. New words are constantly being created as well as new senses are assigned to existing words.

III. PROPOSED SYSTEM

The project proposes an automatic method to estimate the semantic similarity between words or entities using web search engines. Because of the vastly numerous documents and the high growth rate of the web, it is time consuming to analyze each document separately.

Web search engines provide an efficient interface to this vast information. Page counts and snippets are two useful information sources provided by most web search engines. Page count of a query is an estimate of the number of pages that contain the query words. In general, page count may not necessarily be equal to the word frequency because the queried word might appear many times on one page.

So the project proposes a method that considers both page counts and lexical syntactic patterns extracted from snippets and shows experimentally to overcome the above mentioned problems.

The new system presents an automatically extracted lexical syntactic patterns-based approach to compute the semantic similarity between words or entities using text snippets retrieved from a web search engine.

The new system proposes a lexical pattern extraction algorithm that considers word subsequences in text snippets. Moreover, the extracted set of patterns is clustered to identify the different patterns that describe the same semantic relation.

IV. MODULE DESCRIPTION

A module description provides detailed information about the module and its sub modules, which is accessible in different ways within the project. The description is available by reading directly, by generating a short description, or by making an environment check for supported components to check if all needed types and services are available in the environment where they will be used. This environment check could take place during registration/installation or during a separate consistency check for a component.

1. SEMANTIC SIMILARITY
2. CO-OCCURENCES MEASURES
3. LEXICAL PATTERN EXTRACTION
4. LEXICAL PATTERN CLUSTERING

1) SEMANTIC SIMILARITY

Page Count based Co-Occurrence Measures

In this module, word1 and word2 are keyed in. The words are combined and displayed as word pair. The ‘webdocument’ folder located in root folder of the application contains HTML pages. The pages are searched with these words. Three list boxes are provided. The first listbox is populated with the page names containing the ‘word1’. The second listbox is populated with the page names containing the

'word2'. The third listbox is populated with the page names contain both the words. The counts of word1 pages, word2 pages and both words are also displayed in label controls. The values are stored in 'GlobalClass' class and used in successive modules.

2) CO-OCCURRENCE MEASURES

Web Jaccard

In this module, the H(P) page count with word1, H(Q) page count with word2, H(P^Q) page count with word pair are displayed in label controls and Web Jaccard Value is calculated and displayed in a label control.

Web Overlap

In this module, the H(P) page count with word1, H(Q) page count with word2, H(P^Q) page count with word pair, minimum of H(P) and H(Q) are displayed in label controls and Web Overlap Value is calculated and displayed in a label control

Web Dice

In this module, the H(P) page count with word1, H(Q) page count with word2, H(P^Q) page count with word pair, 2 * H(P^Q) are displayed in label controls and Web Dice Value is calculated and displayed in a label control.

Web PMI (Point wise Mutual Information)

In this module, the H(P) page count with word1, H(Q) page count with word2, H(P^Q) page count with word pair, H(P)/N, H(Q)/N, H(P^Q)/N are displayed in labelcontrols and Web PMI Value is calculated and displayed in a label control. In the sample values, 'N' is taken as 10. In real time the 'N' will be 10 to the power of 10 or more.

3) LEXICAL PATTERN EXTRACTION

Search Pattern input with multiple words

In this module, the search pattern is entered in which the first word and last word are taken. In the web pages, the phrase is checked such that the pattern is first word, any number of words and the last word. During the pattern extraction, the skip count number of words can be discarded in the phrase found in the web pages.

Retrieve Pages with given words

In this module, the search pattern is found out from the web pages and the pages names are added in a list.

4) LEXICAL PATTERN CLUSTERING

In this module, the patterns can be clustered using the lexical pattern clustering algorithm. The patterns are clustered and then the count and co-occurrence of the word can be considered. Based on this the word can be extracted. The cluster can be grouped based on the threshold value entered in textbox control, the words are clustered and the n the results are produced in the listbox control.

V. CONCLUSION

Accurately measuring the semantic similarity between words is an important problem in web mining, information retrieval, and natural language processing. Web mining applications such as, community extraction, relation detection, and entity disambiguation; require the ability to accurately measure the

semantic similarity between concepts or entities. In information retrieval, one of the main problems is to retrieve a set of documents that is semantically related to a given user query.

The project proposed a semantic similarity measure using both page counts and snippets retrieved from a web search engine for two words. Four word co-occurrence measures were computed using page counts. It proposed a lexical pattern extraction algorithm to extract numerous semantic relations that exist between two words. Moreover, a sequential pattern clustering algorithm was proposed to identify different lexical patterns that describe the same semantic relation.

The project proposes an empirical method to estimate semantic similarity using page counts and text snippets retrieved from a web search engine for two words. Specifically, it defines numerous word co-occurrence measures using page counts and integrates those with lexical patterns extracted from text snippets. It showed that the proposed method outperforms various baselines as well as previously proposed web-based semantic similarity measures, achieving a high correlation with human ratings. Moreover, the proposed method improved the community mining task, thereby underlining its usefulness in real-world tasks that include named entities not adequately covered by manually created resources.

REFERENCES

i. Antonio Jimeno-Yepes and Alan R. Aronson “Query expansion for UMLS Metathesaurus disambiguation based on automatic corpus extraction,” 2010
ii. Bartosz Broda and Maciej Piasecki “Semi-supervised Word Sense Disambiguation Based on Weakly Controlled Sense Induction,” 2009
iii. Boshra F. Zopon AL_Bayaty, Dr. Shashank Joshi “Empirical Comparative Study to Supervised Approaches for WSD Problem: Survey,” 2015
iv. Bridget T. McInnes , Mark Stevenson “Determining the difficulty of Word Sense Disambiguation,”.Biomedical Informatics 2014
v. Charnyote Pluempitiwiriyawej , Nick Cercone and Xiangdong “An Lexical acquisition and clustering of word senses to conceptual lexicon construction,” 2009
vi. Clayton Scott and Robert D. Nowak Minimax-Optimal “Classification With Dyadic Decision Trees,” 2006
vii. Ganesh Chandra and Sanjay K.Dwivedi “A Literature Survey on Various Approaches of Word Sense Disambiguation,” 2014
viii. Ganesh Chandra and Sanjay K.Dwivedi “A Literature Survey on Various Approaches of Word Sense Disambiguation,” 2014
ix. Kobus Barnard andMatthew Johnson “Word sense disambiguation with pictures,” 2005
x. S. G. Kolte and S. G. Bhirud “Word Sense Disambiguation using WordNet Domains,” 2008
xi. Ling Che and Yangsen Zhang “Study on Word Sense Disambiguation Knowledge
xii. Base Based on Multi-sources,” 2011
xiii. Miguel Ángel Ríos Gaona andAlexander Gelbukh “Web-based Variant of the Lesk Approach to Word Sense Disambiguation,”2009
xiv. M. Rajani Shree and Dr. Shambhavi B.R “Performance Comparison of Word Sense Disambiguation Approaches for Indian Languages,” 2015
xv. Tamara Mart´ın-Wanton , Rafael Berlanga-Llavori and Antonio Jimeno-Yepes Preliminary results for Biomedical Word Sense Disambiguation based on Semantic Clustering,” 2011
xvi. Tomasz Nykiel And Henryk Rybinski “Word Sense Discovery for Web Information Retrieval,” 2008
xvii. Wei Jan Lee and Edwin Mit “Word Sense Disambiguation By Using Domain
xviii. Knowledge,” 2011
xix. Wessam Ga El-Rab , Osmar R. Zaïane and Mohammad El-Hajj “Unsupervised Graph-based Word Sense Disambiguation of Biomedical Documents,” 2013
xx. Yao-Feng Wang, Yue-Jie Zhang, Zhi-Ting Xu And Tao Zhang “Research On Dual Pattern Of Unsupervised And Supervised Word Sense Disambiguation,” 2006
xxi. Yukiko Sasaki Alam “Lexical-Semantic Representation of the Lexicon for Word Sense Disambiguation and Text Understanding,” 2009
xxii. Zhimao Lu, DongMei Fan and Rubo Zhang “Word Sense Disambiguation Based on Vicarious Words ,” 2008

