A New Approach for Computing Semantic Relatedness with Wikipedia

Xinye Zhang, Xiu Li, Zhijian Ruan
Department of Computer Science and Technology, Tsinghua University, Beijing, China
zhangxinye730@163.com,lixiu@cic.tsinghua.edu.cn, rzj19880219@yahoo.cn

Abstract—Semantic relatedness measures are used in many applications in natural language processing and we propose a Wikipedia-based method to compute it. Unlike existed methods that only focus on a small section of Wikipedia (e.g. info box or hyperlinks), our method makes full use of the rich information contained in the Wikipedia page and could get a higher accuracy within reasonable time. In our method, we first use some special sections (e.g. synonyms and hyponyms) in the Wikipedia page to judge whether two concepts are closely related. If they are not, we then use pattern matching to find whether they are related through usual relatedness (e.g. "a part of", "result in", and "is a member of "). And if the relatedness score hasn't been computed out through former steps, we then use a method which makes some improvement on the famous explicit semantic analysis method to compute the relatedness.

Index Terms - semantic relatedness; Wikipedia-based; semantic kernel;

1. INTRODUCTION

Semantic relatedness indicates how much two concepts are related in a taxonomy by using all relations between them (i.e. hyponymic, hypernymic, meronymic and any kind of functional relations including has-part, is-made-of, is-an-attribute-of, etc.). Semantic relatedness measures are used in many applications in NLP (Natural Language Processing) such as word sense disambiguation, information retrieval, interpretation of noun compounds and spelling correction.[1]

Many researches have been done to compute semantic relatedness making use of manually-built thesaurus (e.g. Wordnet) and web resources (e.g. Wikipedia). With a higher coverage of knowledge and high quality, web resources (e.g. Wikipedia) are regarded as better candidate for computing semantic relatedness between concepts and experiment evaluation in [2].

As an internet resource and a collaborative Wiki-based encyclopedia, Wikipedia has various impressive characteristics such as a huge amount of articles, live updates, a dense link structure, brief link texts and URL identification for concepts. With these characteristics, Wikipedia has become an invaluable resource for computing semantic relatedness. However, Wikipedia has a complicated structure, and most researchers only focused on some sections of it (e.g. hyperlinks information, category information).

Since the structure of Wikipedia page is complicated, and different section provides different information, we propose a new method to compute relatedness making full use of the rich information on the Wikipedia page.

The rest of this paper is organized as follows. Section II describes some related works about methods computing semantic relatedness. Section III makes an introduction to Wikipedia structure. Section IV describes our methodology in detail. Section V describes experimental evaluation result of our method. Finally, we make some conclusions in section VI.

2. RELATED WORK

As a manually built thesaurus, WordNet is widely used in many approaches for computing semantic relatedness. The paper [3] defines a term similarity matrix using WordNet to improve text clustering. Their approach only uses synonyms and hyponyms. It fails to handle polysemy, and breaks multi-word concepts into single terms. The paper [4] proposes to transform the WordNet into a graph and compute semantic relatedness using paths in it.

While WordNet represents a well-structured taxonomy organized in a meaningful way, it also has limitations: it has a small coverage and could not update timely.

As an internet resource and a collaborative Wiki-based encyclopedia, Wikipedia has various impressive characteristics such as a huge amount of articles, live updates, a dense link structure, brief link texts and URL identification for concepts. With these characteristics, Wikipedia has become an invaluable resource for research in various areas such as AI, NLP, Web mining and Semantic Web .[5]

WikiRelate is a method proposed in [1]. Given a pair of words w1 and w2, WikiRelate searches for Wikipedia articles, p1 and p2, which respectively contain w1 and w2 in their titles. Semantic relatedness is then computed using various distance measures between p1 and p2. These measures either rely on the texts of the pages, or path distances within the category hierarchy of Wikipedia.

ESA is another Wikipedia-based method proposed in [2], it is a method that represents the meaning of texts in a high-dimensional space of concepts derived from Wikipedia. Assessing the relatedness of texts in this space amounts to comparing the corresponding vectors using conventional metrics (e.g., cosine). Compared with the previous algorithms, using ESA results in substantial improvements in correlation of computed relatedness scores with human judgments: from r =0.56 to 0.75 for individual words and from r =0.60 to 0 .72 for texts.

As we can see, most existed methods only focused on a part section of Wikipedia (e.g. category information), and the rich information on the Wikipedia page is not fully utilized. So we propose a method for computing semantic relatedness making full use of the rich information on the Wikipedia page.

3. INTRODUCTION TO WIKIPEDIA

Wikipedia was launched in 2001 with the goal of building free encyclopedias in all languages. Today it outstrips all other encyclopedias in size and coverage, and is one of the most visited sites on the web. Out of more than three million articles in more than 200 different languages, one-third is in English, yielding an encyclopedia almost ten times as big as the Encyclopedia Britannica, its closest rival. Since Wikipedia's high quantity and high quality of articles, it has been used in many areas, such as IR, NLP and web Mining.

The structure of Wikipedia is complicated and it has many kinds of pages (e.g. category pages, image pages, template pages, talk pages) and hyperlinks (e.g. redirect hyperlinks, out-category hyperlinks, common hyperlinks). In our work, category pages, common article pages and all kinds of hyperlinks are the focus we concentrate on.

Category pages are used to provide category information, including sub-categories and parent categories. All category information of Wikipedia results in a tree-like directed graph and Fig. 1 shows a section of the whole directed graph.

Wikipedia contains only one article for any given concept (called preferred term). There are several kinds of hyperlinks on a common article page: redirect hyperlinks exist to group equivalent concepts with the preferred one ("Operating System" may also be called "OS"), out-category hyperlinks indicate which categories this article belongs to, common hyperlinks are the manually tagged concepts that exist in the whole Wikipedia.

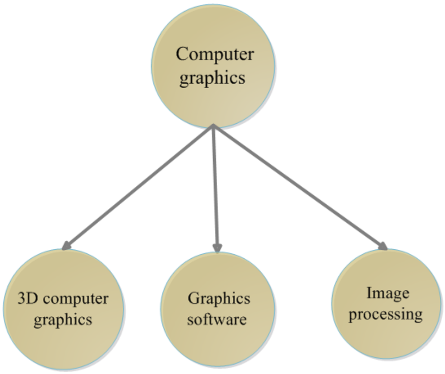


Figure 1 a subgraph of Wikipedia

Disambiguation pages are provided for polysemous concepts. A disambiguation page lists all possible meanings associated with the corresponding concept, where each meaning is discussed in common article. For example, the disambiguation page for "OS" lists more than thirty meanings of this term.

4. METHODOLOGY

Since Wikipedia article page contains rich information, we should make full use of these valuable resources. First, we need to solve the problem of disambiguation. Given a pair of terms, we should identify the proper meaning of each term, so we could get more accurate semantic relatedness of them in the following steps. Second, we design a method for computing the semantic relatedness of terms making full use of the rich information of Wikipedia. Evaluation is then done on the experiment results.

A. Disambiguation

Since the existence of polysemous terms, we need to find the proper meaning for these terms. If one term is given in sentence or paragraph, we then use the similarities between the sentence and every wiki article, finally the one with highest score would be considered as the proper meaning of this term. [2] If only one single term is given, then we have two options: either we choose the most common meaning (i.e. the first meaning) as the term's meaning or we choose the meaning pair that could produce the highest semantic relatedness. Here we opt to regard the most common meaning as the proper meaning of the term.

B. Computing Semantic Relatedness

The structure of Wikipedia page is complicated, and different section plays different role in computing semantic relatedness. We first find the closely related terms making use of some manually-tagged information in the Wikipedia page, and then we use a method to compute the relatedness between term pairs whose relatedness are not manually tagged out. Steps for computing semantic relatedness are shown in Fig. 2.

```
Input: a term pair A and B
Output: the semantic relatedness score for A and B
Algorithms:
1. Find the proper meaning of A and B
2. Judge whether A and B are closely related through
   some manually-tagged information, if they are,
   then just give them a high score and computation
   is over, otherwise continue to do the following
   steps.
3. Compute the relatedness of A and B using a
   pattern matching, if they are related through usual
   relations, then give them a high score, otherwise
   use an improved ESA to compute the relatedness
   score between them.
```

Figure 2 steps for computing semantic relatedness

1) Get closely related terms through manually-tagged information

On a Wikipedia page, many closely related terms have been manually tagged out in several sections, and these sections mainly contain synonyms, see also section, info box and category information. If two terms are closely related and have been tagged out in these sections, what we need to do is giving them a high score and no further computation is needed, which will result in a reduced computation and high accuracy. However, if they are not found in the manually-tagged sections, we then need to do further computation to get their semantic relatedness.

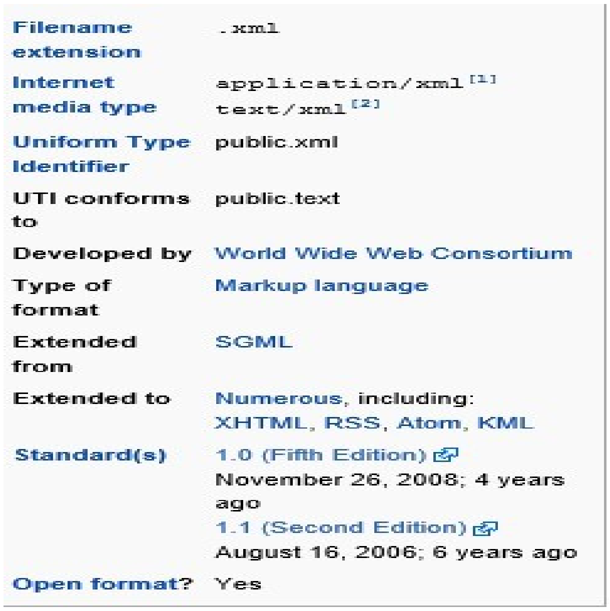
a) Synonyms

Since the existence of synonyms, one concept may correspond to several terms (e.g. USA, US, America, United States and United States of America correspond to the same concept). There is no doubt that the relatedness between one term and its synonyms would be absolutely high (we set the relatedness as the full score 10). In a common Wikipedia page, the term in bold appeared in first paragraph and the redirect links show the synonyms of one term, and we could easily get them. So given a term pair, we first check out whether they are synonyms, and if they are, then there's no need to do following computation, otherwise we need to do following checks to compute the final semantic relatedness between them.

b) See Also

The section of "See Also" lists some related concepts of the article, and the relatedness between the article title and these concepts are often close (e.g. the concepts "XML Protocol", "WBXML", " Binary XML" all appear in the "See Also" section of the concept "XML"). What's more, for some hyperlinks in the "See Also" section, we need to follow the link and check the page content of these links (e.g. "List of XML markup languages" appeared in the "See Also" of "XML"). Luckily, these links has some common features (e.g. they often contain the word "List" or "Outline") which we could use to identify them. For those hyperlinks appeared in the "See Also" section, we also should give a high score (experiment shows that a point of 8 is reasonable).

c) Info Box



Filename extension	.xml
Internet media type	application/xml [1] text/xml [2]
Uniform Type Identifier	public.xml
UTI conforms to	public.text
Developed by	World Wide Web Consortium
Type of format	Markup language
Extended from	SGML
Extended to	Numerous, including: XHTML, RSS, Atom, KML
Standard(s)	1.0 (Fifth Edition) November 26, 2008; 4 years ago 1.1 (Second Edition) August 16, 2006; 6 years ago
Open format?	Yes

Figure 3 a section of one info box

One example for info box is shown in Fig. 3. Info box plays an important role in the Wikipedia article page, and it provides some information about the features of the corresponding concept. So if one concept appears in another concept's info box, then their relatedness is very close (experiment shows that a point of 8 is reasonable). However, not every Wikipedia article has info box, so finding closely related concepts through info box is limited.

d) Category Information

Category information in common Wikipedia page provides important information in the Wikipedia page, and it shows which categories one concept belongs to. What's more, category information in a category Wikipedia page shows one category's parent categories and its child categories. Category information is widely used to compute the relatedness of term pair and most methods are based on the shortest path distance between two concepts. Since Wikipedia is a densely connected directed graph, there would be at least one path between any concept pair, so only small path distance could provide useful information. In the directed Wikipedia graph, two concepts with the shortest distance 2 means that they have parent-child relation, so their relatedness score should be high (experiment shows that a point of 8.3 is reasonable). For those with shortest path distance 2, their situation could be high different. For example, the relatedness score of "car insurance" and "car safety" should be high, while the relatedness score of "car insurance" and "car body styles" should be low. So computing the semantic relatedness score with shortest path distance could have high deviation, and this is why we only choose to compute those with shortest path distance 1.

2) Compute semantic relatedness through text article content

If the semantic relatedness of a term pair hasn't been computed in the former steps, then we should check the article content to compute the score. First we use the pattern matching method to judge whether two concepts are closely relatedness through usual relations (e.g. the relation of "a part of", "result in", "is a member of "). If they are not related in usual relation, then we use statistics method to compute the relatedness.

a) Compute relatedness with pattern matching

In most case, two concepts are closely related because there is at least a usual relation between them. For example, "the sun" makes the weather "warm", a "cup" could be used to drink "water" and "vehicle gas" contributes to "global warming". So we could use pattern matching to find those relations. Given a concept pair A and B, we first get their corresponding Wikipedia article content, and then we do word segmentation making use of Wikipedia thesaurus. We download the xml thesaurus of Wikipedia and get the concept list through xml parser. With this gotten concept list, we could identify the proper concepts existed in the Wikipedia and keep those multi-word concept unbroken. Then we try to find whether the two concepts are related through defined usual relations. During the process of pattern matching, not only the two concepts A and B are considered, but also their synonyms are under consideration. Since the Wikipedia page content is limited, we also make use of search engine results to help find those usual relations.

If two concepts are found related through usual relations, then we just need to give them a uniform score (we give 8.5 as the relatedness score). Otherwise we need to do following checks.

b) Compute relatedness through statistics

If the relatedness score hasn't been computed out through former steps, then we use a method which is similar to the one in the paper [2] to compute it. Here we make an improvement: when compute the weight of different terms, we multiply a coefficient (1.25) to those manually-tagged terms (e.g. hyperlinks and terms in bold).

In a common article, it is natural for us to regard those manually-tagged terms more important than those not tagged.

What's more, terms in different structure can also mean different importance. For example, the first three paragraphs of a Wikipedia article are usually used to explain one concept and they provide more important information than those paragraphs after them. So after we get the weight through TF-IDF scheme, we could multiply a coefficient to the gotten weight of terms which are more important and our experiment shows that we could get a better result.

5. Evaluation

To assess semantic relatedness, we use the famous WordSimilarity-353 collection which contains 353 word pairs. Each pair has 13–16 human judgments, which were averaged for each pair to produce a single relatedness score. Spearman rank-order correlation coefficient was used to compare computed relatedness scores with human judgments.

Table 1 Comparasion of different algorithms

Algorithms	Correlation with humans
WikiRelate!	0.19-0.48
ESA-Wikipedia	0.75
Ours	0.81

As shown in Table 1, our approach has obvious improvement compared to traditional approaches.

6. Conclusion

In this paper, we points out the disadvantage of traditional methods for computing semantic relatedness and we propose a new method to compute it. In our method, we make full use of the manually-tagged information in the Wikipedia page and use pattern matching to mine the usual relatedness between concept pair and ESA-like methods to compute the relatedness of terms that are not obviously closely related.

Our future work will concern on further applications based on semantic relatedness such as text classification and resource recommendation.

REFERENCES

[1] Michael Strube and Simon Paolo Ponzetto. WikiRelate! Computing semantic relatedness using Wikipedia. InAAAI'06, Boston, MA, 2006.

[2] Gabrilovich, E. and S. Markovitch. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis, Proc. of IJCAI'07I.

[3] L. Jing, L. Zhou, M. K. Ng, and J. Z. Huang. Ontology-based distance measure for text clustering. In Workshop on Text Mining, SIAM International Conference on Data Mining, Bethesda, MD, 2006.SIAMR.

[4] Hirst, G. & D. St-Onge (1998). Lexical chains as representations of context for the detection and correction of malapropisms. In C. Fellbaum (Ed.),WordNet: An Electronic Lexical Database, pp. 305–332. Cambridge, Mass.: MIT Press

[5] K. Nakayama, T. Hara, and S. Nishio, "Wikipedia mining for an association web thesaurus construction," in Proc. of IEEE International Conference on Web Information Systems Engineering (WISE 2007), pp. 322–334,2007

[6] L. Denoyer and P. Gallinari. The Wikipedia XML Corpus.SIGIR Forum, 2006.

[7] E. Gabrilovich and S. Markovitch. Overcoming thebrittleness bottleneck using wikipedia: enhancing text categorization with encyclopedic knowledge. In National Conference on Artificial Intelligence (AAAI),Boston,Massachusetts,2006.

[8] Mehran Sahami and Timothy Heil-man. A web-based kernel function for measuring the similarity of short text snippets. InWWW'06. ACM Press,2006.