A specific word relatedness computation algorithm for news corpus

1, Liu Jinpan 2, He Liang 3, Lin Xin 4, Xu Mingmin

Department of Science and Technology, East China Normal University
{jpliu ,lhe9191 , xlin , mmxu}@ica.stc.sh.cn

Abstract - The determination of relevance of any given word combination is a profound issue in many applications of natural language processing (NLP) [1]. In this paper, a new method is proposed to compute the relativity of terms in the news corpus. And for the reason that each single news belongs to a specific channel, the news corpus should be divided into several sub-corporuses according to different categories. Secondly, the feature of news document is exploited, as a matter of fact, the co-occurrence in the title of the news and that in the news content are attached with different weights. At last, we introduce the Wikipedia corpus to overcome some innate shortcomings of the news corpus.

Keywords-component; term co-occurrence; news corpus; word relatedness; Wikipedia;

I. INTRODUCTION

Word relatedness measurement is a basic research topic in the fields of nature language processing, such as word sense disambiguation [2], information retrieval [3], information extraction pattern induction [4], and paraphrase detection [5]. Word relatedness is the measurement of the relationship between two terms in the scale of 0 to 1. Word similarity and word relatedness are two tangled concepts [6]. Take the two words “Rocket” and “NBA” for illumination, the similarity of them is very little, meanwhile the relatedness between them is comparatively very large. As a matter of fact, the word similarity represents the same meaning of two terms while the relatedness indicates that a relationship exists between them. Meanwhile, it is more likely that two words tend to have the closer correlativity if the two have a basically large similarity. It is important to note that semantic relatedness is a more general concept than similarity; similarity entities are semantic related by virtue of their similarity, but dissimilar entities may also be semantic related by lexical relationships which are required more in Computational applications.

There are two main methods to compute the term relatedness, one is to use a dictionary such as WordNet (http://wordnet.princeton.edu/), HowNet (http://www.keenage.com) or domain ontology such as medical subject Headings (MeSH http://www.nlm.nih.gov/mesh/). The other method is to use the corpus. The ontology specifies the relationships between different concepts. Each concept possesses several primitives which can be used for computing the correlativity between terms. The commonly appreciated ontology delicately created by the experts can use little space to present complicated relationships between terms. Even so, the drawback of ontology cannot be pathologically ignored. The relationship in the ontology is subjective based on personal understanding and also insensitive meaning that it can’t evolve with time and is hard to absorb new words. The other method to compute correlativity is using the corpus, which avoids the aforementioned drawbacks caused by the ontology for the corpus comes from the real life and could update in time. The basic methodology to compute term relatedness using corpus is counting the co-occurrence, the more frequently two terms co-occurrence in the corpus, the closer relationship they have. To compute the co-occurrence of any two given words, we define a “window” which restrains a maximum distance between two-co-occurrence-terms. The window size parameter allow us to look at different scales smaller window size will identify fixed expressions (idioms as bread and butter) and other relations that hold over short ranges; large window sizes will highlight semantic concepts and other relationships that hold over large scales. In our experiments, a window is defined as the title of a news document or the content of it.

The remainder of the paper is organized as follows. In section 2, we survey the related work including semantic similarity and word relatedness measurements. A detailed description of the method is presented in Section3. The analysis of experimental results is in Section4. In Section5, the conclusion and future work is drawn.

II. RELATED WORK

Semantic correlativity measurements have been used in applications related to Semantic Web, such as automatic annotation of Web pages [8], keyword extraction for inter-entity relation representation. Research in the area of computing semantic correlativity and similarity can be generally divided into two categories. One depends on the structured lexical databases such as WordNet and the other is built on the large text corpus [9].

The first time using lexical databases to measure semantic similarity can be found in [10]. The straightforward method they adopt to calculate similarity between two words (concepts) is to find the length of the shortest path connecting the two words in the lexical databases. A
frequently acknowledged problem with this approach is that it relies on the notion that all links in the taxonomy represent a uniform distance. Research in this area proceeded in two directions. Firstly new lexical databases were introduced in. Early on, WordNet was used to provide a broad coverage lexical database, for example in [11], [12]. Later Jarmssz and Szpakocicz(2003)explored the use of Roget’s Thesaurus for computing semantic similarity. Secondly major advances were achieved by developing more sophisticated of semantic similarity and relatedness.

An objective measure based on the information theoretic notion of mutual information to compute the word association was proposed in [3]. According to Fano(1961), if two points(words), \(x\) and \(y\) have probabilities \(P(x)\) and \(P(y)\), then their mutual information \(I(x, y)\), is defined to be

\[
I(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)}
\]

If there is a genuine association between \(x\) and \(y\) then the joint probability \(P(x, y)\) will be much larger than chance \(P(x)P(y)\), and consequently \(I(x, y) \geq 0\). If there is no interesting relationship between \(x\) and \(y\) then \(P(x, y) \approx P(x)P(y)\), and thus, \(I(x, y) \approx 0\). If \(x\) and \(y\) are in complementary distribution, then \(P(x, y)\) will be much less than \(P(x)P(y)\), forcing \(I(x, y) \leq 0\). This method makes use the co-occurrence information to compute the relatedness.

III. METHOD

3.1 Procedure

We propose a method which makes use the feature of news and integrates both the term co-occurrence and the web corpus. in section 2, we describe the feature of news and the construction of the corpus. we then describe how to compute the term co-occurrence in section 3. Following we introduce the Wiki relatedness, the news corpus includes a lot of news document which have a high speed of update while some relation between terms as is known to all does not include. Then we integrate all the things talked about above and give a method to compute the term correlativity.

3.2 The feature of news corpus and news division

As we know that News different from other kinds of documents, has its own characteristics.

(1) The news corpus can be divided into several sub-corporuses according to different categories.

(2) High update frequency.

(3) The co-occurrence in the title of the news and that in the news text should be treated differently.

(4) The large amount of new emerging concepts.

The same word in different channels has different meanings, for example “rocket” in the sports channel means a NBA team, while in science channel it means a spacecraft or aircraft, and it represents a missile in military channels. So we need compute the correlativity in different channels. “rocket” in the sports channels may have a large co-occurrence with NBA, while in the science channel it may appear more with the space.

As we know that the title represents the theme of the news document. So the co-occurrence in the title should be attached with a bigger weight than that appears in the news content. The structure of the corpus is shown in figure 1

![Figure 1. the structure of the corpus](image)

3.3 The Model

Then we introduce some mathematical notations. Let \(DT\) denote an occurrence matrix of order \(m \times n\). The columns of \(DT\) correspond with the term denoted by 1 to \(n\). The rows of \(DT\) usually correspond with news denoted by 1 to \(m\). one assumption that we make is that \(DT\) is a binary matrix, that is, each element of \(DT\) equals either zero or one. Let \(DT^T_i\) denote the element in the kth row and ith column of \(DT\). \(DT^T_i\) equals one if word i occurs in the news that corresponds with the kth row of \(DT\), and it equals zero otherwise. Let \(C\) denote the co-occurrence matrix of the term \(1 \cdots n\). \(C\) is a symmetric non-negative matrix of order \(n \times n\). Let \(C^T_{ij}\) denote the element in the ith row and jth column of \(C\). \(C^T_{ij}\) equals the number of co-occurrences of word i and j. It follows from this that

\[
C = DT^T DT
\]

Where \(DT^T\) denotes the transpose of \(DT\). Moreover, the assumption that \(DT\) is a binary matrix implies that \(C\) is an integer matrix.
Let $S_j$ denote either the total number of occurrences of object $i$.

$$S_j = C_{ij} = \sum_{k=1}^{m} DT_{ki}$$

(2)

### 3.4 The term co-occurrence computation

There are multiple measures for co-occurrence, such as Association Strength Coefficient, Inclusion Coefficient, Jaccard Coefficient, and Cosine coefficient, formulated in Table 1. All are applied generally in Information Retrieval, Information Extraction, and entity relation identification. The measurement can be divided into two fundamentally different types. One type is called set-theoretic similarity measures, which can be interpreted as measures of the relative overlap of two sets. The other type is called probabilistic similarity measures, which can be interpreted as measures of the deviation of observed co-occurrence frequencies from expected co-occurrence frequencies under an independence assumption [13]. The cosine, the inclusion index, and the Jaccard index are examples of set-theoretic similarity measures, while the association strength is an example of a probabilistic similarity measure.

<table>
<thead>
<tr>
<th>Measure Name</th>
<th>Measure Formulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Association Strength Coefficient</td>
<td>$C_{ij} / S_j S_j$</td>
</tr>
<tr>
<td>Inclusion Coefficient</td>
<td>$C_{ij} / \min(S_i, S_j)$</td>
</tr>
<tr>
<td>Jaccard Coefficient</td>
<td>$C_{ij} / S_j + S_j - C_{ij}$</td>
</tr>
<tr>
<td>Cosine Coefficient</td>
<td>$C_{ij} / \sqrt{S_j S_j}$</td>
</tr>
</tbody>
</table>

We use the generalized similarity Coefficient proposed by IDO DANGN which is define as

$$SIG = \frac{2^{1/p} C_{ij}}{(S_i^p + S_j^p)^{1/p}}$$

to compute the content co-occurrence. The value of $p$ is in $R$ . For all values of the parameter $p$, the generalized similarity index takes values between zero and one. The reason why we use this equation is that when the $p$ changes the generalized similarity could present all the measurements for co-occurrence except the Association Strength Coefficient. It can maximize the possibility of all

It is not difficult to see that

$$\frac{C_{ij}}{S_j} \leq \frac{C_{ij}}{S_i + S_j - C_{ij}} \leq \frac{C_{ij}}{\sqrt{S_j S_j}}$$

As the length of the title is short, term A or term B appearing in the title is frequency while term A and term B appearing in the title is seldom. In order to obtain a relative high co-occurrence value of A and B, it is possible to use the Inclusion Coefficient to compute the title co-occurrence.

### 3.5 The Wiki relatedness

Wikipedia is a website that uses wiki software, allowing the easy creation and editing of any number of interlinked web pages to create an encyclopedia. Being a collaborative open source medium, it is highly structured.

The news corpus has high update frequency. Most portal sites provide 24-hour comprehensive and instant information. While a lot of relations between terms as known to all will always be ignored. For example apple and fruit have a great relevance but they don’t appear together in the news corpus frequently. In this situation, we introduce the Wikipedia. Wikipedia provides a semantic network for computing semantic relatedness in a more structured fashion than a search engine and with more coverage than WordNet. [14]In the Wikipedia, each webpage has a theme and a lot of marking fonts which is consider a concept with a Hyper Link to its own page. The marking fonts included in the theme’s page must have some relation to the theme. Intricate relationships between concepts in Wiki form a complex network. Each term in the database has a list of related terms and they have theirs. The relationship can be shown in the figure 2.

![Figure 2. The structure of the web corpus](image)

Based on the hierarchy of relationship between terms, relationship are divided into different types, as we know that the fewer levels, the closer relationship between them, the rocket has a greater relevance with Houston Rocket, than it with NBA. In the network, we only use two hierarchies of the network to reduce the complexity of calculation. If two terms are direct connect in Wiki entity network, the Wiki relatedness is A. when the distance between two terms is 2, the Wiki relatedness is B. In this paper we define the value of A is 1 and the value of B is 0.5.

### 3.6 The method we propose

As we mentioned above, the method we propose is an integration of the co-occurrence in the title, the co-occurrence in the news content and the Wikipedia.
relatedness, the news we use to compute relevance has already been divided into different channels.

\[
\alpha \frac{2\sqrt[p]{c_y}}{(s^p_i + s^p_j)^{1/p}} + \beta \frac{c_y}{\min(S_i, S_j)} + (1 - \alpha - \beta)w_j \quad (i = 0, 1, 2,)
\]

In this equation \(W\) means the weight of the Wiki relatedness. the \(i\) present the hierarchy at which the terms have relation, if they don’t have any relation in the database the hierarchy is 0. The co-occurrence in the title we use the Cosine Coefficient to measure. The co-occurrence of words in the news text we use the generalized similarity Coefficient because it could present all the set-theoretic similarity measures, and the value we could get from our experiment.

IV. EXPERIMENT

4.1 Dataset Analysis

Based on the features of news, we construct a news corpus though crawling the Chinese portal website such as Sohu (http://www.sohu.com) every five minutes for one year. After preprocessing each news, we have id, title, time, url, text in the database, then we segment the text of the news. The corpus is divided into different channels, for example auto channel, sports channel, health channel, news channel. Each news in the corpus includes title and content. In this paper we use three channels. Table 2 shows us the detail information of our corpus.

<table>
<thead>
<tr>
<th>Channels</th>
<th>News items</th>
<th>Terms in title</th>
<th>Terms in content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sports</td>
<td>439149</td>
<td>9737462</td>
<td>23859312</td>
</tr>
<tr>
<td>News</td>
<td>456673</td>
<td>13852024</td>
<td>27363467</td>
</tr>
<tr>
<td>Military</td>
<td>14041</td>
<td>371428</td>
<td>6963946</td>
</tr>
</tbody>
</table>

Term in title means the number of terms pair in one channel. And News items means the number of news in one channel. To compute the title co-occurrence,, we define the title as a basket, if two terms appear in the same title we consider they co-occurrence once, at the same time we record the frequency of each term. The same situation in the content.

4.2. Experiment Result

Result 1: The different frequency in each channel

Figure 3 demonstrates that each “rocket” represents a specific frequency in each channel—”Sports”, “Military” and “News”. The ratio is obtained by dividing the number of the appearance of word “rocket” with the number of the document.

In the following experiments, we random select ten words of which five words have greater relevance with rocket, five words have little. We make a comparison of the parameter value and decide the better one.

**Result 2: The value of \(p\) between 1 and 50 is better**

In picture 2, \(\alpha\) and \(\beta\) are set to a fixed value and we give a different value to \(p\). The photograph shows that when the value of \(p\) is 1, it has a greater effect in computing the relatedness of term which has closer relationship with rocket, and if the value of \(p\) is 50 it is good at computing the relatedness of term which has little relationship with rocket. Then we think that the value of \(p\) between 1 and 50 is better.

**Result 3: The weight of Wiki relatedness is 0.4**

In order to reflect the different weight of title and content co-occurrence, we set the weight of title is twice of content. The value of \(p\) is 5 and the weight of Wiki relatedness is different, we can see that when the value of weight is 0.6 it has a good effect both in computing the relatedness of closer terms and remoter terms. If the weight of Wiki relatedness is 0.6, it has a great influence on the
relatedness computing, a tern has a direct relationship in Wiki and the value of Wiki relatedness is 0.6 and that can’t reflect the influence of co-occurrence so we think that the weight of Wiki relatedness is 0.4.

Result 4: The algorithm we proposed has a perfect result in computing the term relatedness

In all we selected 10 words, we use the algorithm we proposed and take the value of p is 5, the weight of Wiki relatedness is 0.4. We find that it is good at computing the relatedness both in two situations. The result is shown in figure 6

<table>
<thead>
<tr>
<th>Term</th>
<th>Value</th>
<th>Term</th>
<th>Value</th>
<th>Term</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>NBA</td>
<td>0.454</td>
<td>俄罗斯</td>
<td>0.485</td>
<td>NBA</td>
<td>0.430</td>
</tr>
<tr>
<td>篮球</td>
<td>0.451</td>
<td>运载</td>
<td>0.477</td>
<td>美国</td>
<td>0.424</td>
</tr>
<tr>
<td>西部</td>
<td>0.447</td>
<td>卫星</td>
<td>0.453</td>
<td>航天</td>
<td>0.421</td>
</tr>
<tr>
<td>黄蜂</td>
<td>0.327</td>
<td>航空</td>
<td>0.443</td>
<td>飞机</td>
<td>0.418</td>
</tr>
<tr>
<td>球队</td>
<td>0.300</td>
<td>武器</td>
<td>0.433</td>
<td>俄罗斯</td>
<td>0.408</td>
</tr>
<tr>
<td>马刺</td>
<td>0.294</td>
<td>飞行</td>
<td>0.428</td>
<td>国际</td>
<td>0.407</td>
</tr>
<tr>
<td>赛季</td>
<td>0.280</td>
<td>国防</td>
<td>0.423</td>
<td>载荷</td>
<td>0.248</td>
</tr>
<tr>
<td>连胜</td>
<td>0.275</td>
<td>美国</td>
<td>0.422</td>
<td>球队</td>
<td>0.243</td>
</tr>
<tr>
<td>球员</td>
<td>0.262</td>
<td>国际</td>
<td>0.418</td>
<td>姚明</td>
<td>0.238</td>
</tr>
<tr>
<td>美国</td>
<td>0.243</td>
<td>雷达</td>
<td>0.231</td>
<td>北京</td>
<td>0.202</td>
</tr>
</tbody>
</table>

From the table we can see that the same term in different channels have different relevance with rocket, for example Yao in Sports channel have very close relation with Rocket but in News channel it doesn’t have the close relation. As we know the Sports channel pay more attention to the basketball while the News channel is more concern about the people’s daily life. In our experiment we find that the wiki have very important impact on the relevance computing, and this is reasonable to expect. The relation in the Wikipedia is commonly accepted.

V. CONCLUSION AND FUTURE WORK

In this paper, a calculation of terms relatedness in news corpus is presented. In the experiment, the algorithm give the explanation of the parameter value explanation and gets a better result, which fits human’s judgment. In the future work, we are attempting to integrate the co-occurrence and the ontology methods to make use the feature of corpus and overcome the shortcomings of general ontology.

VI. REFERENCES


