# A New Semantic Similarity Measuring Method Based on Web Search Engines

GANG LU[1*] PENG HUANG[2] , LIJUN HE[3], CHANGYONG CU[4] AND XIAOBO LI[5]

[1]School of Information
Zhejiang University of Finance & Economics
18 Xueyuan Street, Hangzhou 310018, CHINA
hz_lugang@163.com

[2]Zhejiang University
1 Yugu road, Hangzhou 310018, CHINA
huangpeng1126@gmail.com

[3]School of Information
Zhejiang University of Finance & Economics
18 Xueyuan Street, Hangzhou 310018, CHINA
helj@zufe.edu.cn

[4]School of Mechanical Engineering
Hangzhou Dianzi University
Xiasha, Hangzhou 310018, CHINA
chuyangyong@pmail.ntu.edu.sg

[5]School of Information
Zhejiang Education Institute
140 Wensan Road, Hangzhou 310012, CHINA
oboaixil@126.com

***Abstract:*** -
Word semantic similarity measurement is a basic research area in the fields of natural language processing, intelligent retrieval, document clustering, document classification, automatic question answering, word sense disambiguation, machine translation, etc.. To address the issues existing in current approaches to word semantic similarity measurement, such as the low term coverage and difficult update, a novel word semantic similarity measurement method based on web search engines is proposed, which exploits the information, including page count and snippets, in retrieved results to do calculation. The proposed method can resolve the issues mentioned above due to the huge volumes of information in the Web. The experimental results demonstrate the effectiveness of the proposed methods.

*Key words*: Web Search Engine; Semantic Similarity; Intelligent Retrieval; Word semantics; WordNet;

## 1 Introduction

Word semantic similarity measurement is a method on how to calculate or compare the semantic similarity between the two words. As a basic research on semantic understanding in nature language handling field, word semantic similarity measurement is attracting more and more researchers' attention. It plays an important role in underlying many higer level applications, and even becomes an important step or a key point in some research fields.

It's simple for the human being to judge whether two words are similar. For example, given word pairs of "honey-bee" and "paper-car", an adult can easily conclude that there's more similarity within word pairs "honey-bee" than "paper-car". However, for a

[*] Corresponding author

computer, it is an exactly difficult task, which involves philosophy, psychology, cognitive science, artificial intelligence and other fields of knowledge. Many algorithms and theories about word semantic similarity measurement have been proposed and improved to promote the judgment between words. So far, the proposed technologies[1-10] can be put forward and divided into four categories: machine-readable dictionary based approaches, knowledge-based approaches, semantic network-based approaches and corpus-based statistical approaches. The precision of the first two approaches is low due to their technology lag and resource restrictions. The last two approaches are the main steams today, and most semantic network-based methods adopt Princeton university's WordNet[11] semantic dictionary.

Semantic network can be seen as a collection of interconnected nodes, where nodes represent the concepts and the lines connecting the nodes represent all kinds of relations between the concepts, such as synonyms relation, antisense relation, and etc[12]. In the field of natural language processing, the most popular semantic network is WordNet[11]. To get the lexical semantic information, WordNet uses semantic attributes to organize the dictionary. All terms in WordNet are organized according to four main categories: nouns, verbs, adjectives and adverbs, and there exist various semantic relations between these terms. Many semantic similarity algorithms based on WordNet have been proposed, such as algorithms based on WordNet semantic network path[4] [13] [14] and algorithms based on information theory[1] [15] [2].

Vocabulary similarity computation based on corpus statistics is an empirical method, which calculates the vocabulary similarity on the basis of observable language fact, not just dependant on the linguist's intuition. Usually this kind of word similarity is known as the word distribution similarity (corresponding to word semantic similarity phase), which is based on the assumption that the two words are similar in the similar context. With large-scale corpus, the statistical information of vocabulary's context is used as a reference for semantic similarity calculation[16, 17]. The researches of vocabulary similarity based on corpus statistics mostly adopt the method of context statistical description, e.g. based on such a conclusion “The context of a word can provide enough information for the word's definition”. The context of a word is usually defined as the words around them within a certain window of which size is usually set as 2. To be distinguished from the WordNet-based semantic similarity, the vocabulary similarity obtained from the corpus is referred as "statistical similarity" or "distribution of similarity."
To some extent, WordNet can be seen as a semantic dictionary, which establishes various semantic relations between vocabularies, including synonyms, antisense, context, or part-whole relations. Many researchers have made an enormous progress in adopting WordNet to calculate lexical semantic similarity. However, the following disadvantages exist in the WordNet based methods: the establishment of a semantic dictionary requires experts and takes a lot of manpower and time; after establishing, it is difficult to update, especially to reflect the new vocabulary and language in time. In addition, the dictionary is greatly affected by the compilers' subjective effects, and sometimes fails to reflect the objective reality. To solve the problems, corpus is proposed to calculate the vocabulary similarity. It reflects the correlation between vocabularies in a more objective manner than the semantic dictionary does. And it can find effective correlation between the strings which can not be obtained through usual human observation, especially for new vocabularies. In recent years, with the rapid development of hardware technology, setting up a large-scale corpus becomes a reality. And due to the emergence of statistical techniques and their wide applications in data mining and machine learning, corpus-based semantic similarity calculation researches have been developed rapidly. However, corpus-based method has the following disadvantages: performance has to depend more on the quality the corpus, sparse data problem, and susceptible to noise interference.

To address the above problems, this paper will present a new word semantic similarity measurement method, which is based on web search engines. The notable feature is that the whole Web environment is taken as an immense corpus, the functions provided by the existing search engines are used to obtain relevant statistical information, and then the semantic similarity between the vocabularies can be calculated. With the large-scale of online Web environment and the extensiveness of the covered fields, this method can solve the low term coverage problem existing in the semantic similarity measurement method based on the traditional WordNet based dictionary.

Figure 1 shows the steps that are discussed in the proposed methodology. Firstly, the whole Web environment is taken as a corpus for word similarity calculation, and the statistical information can be obtained by search engines (Google in this paper). The low term coverage problem of WordNet and the sparse data problem of Corpus can be alleviated because of the huge volumes of information in the Web. Secondly, after adopting the synonyms defined in WordNet, more information can be retrieved from the search results to support word similarity calculation. Thirdly, the revised feature model is employed in the proposed Snippet category (elaborated later on) to improve the accuracy of Snippet category, and further improve the accuracy of word similarity calculation which is based on the Snippet category.

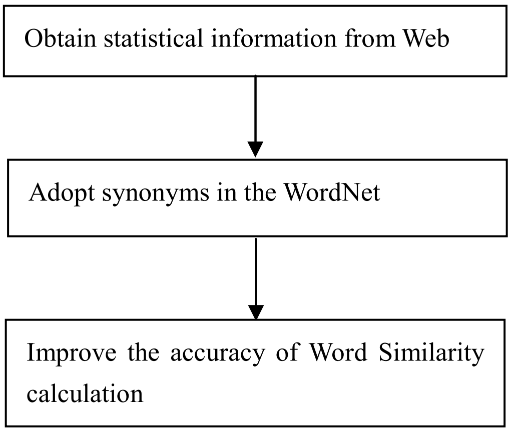

Obtain statistical information from Web
Adopt synonyms in the WordNet
Improve the accuracy of Word Similarity calculation

Figure 1. The basic steps in the proposed new word semantic similarity measurement method

## 2 A Method of Semantic Similarity Measurement Based on Web Search Engines

### 2.1 Relevant Background

Distribution similarity calculation is based on the pre-setup corpuses, which are usually fixed-size, and some of them even area relative. Although the use of corpus can solve the low term coverage problem of the WordNet based methods to a certain extent, this problem still exists, especially with the growing new words in the Web environment. As we all know, the new words and phases in the Web environment emerge or grow dynamically with a high speed everyday.

To address the above problems, more researchers turned to Web-based word similarity calculation, taking the entire Web environment as a corpus. The collection of Web pages is generally a heterogeneous collection. It has two distinct features: massive data and high dynamicity. The massiveness of the data makes the Web document collection have a good vocabulary collection coverage. In fact, it can be approximately treated as 100% coverage. And high dynamicity also makes the Web documents have good new vocabulary coverage.

How to effectively use the large amounts of Web data in the information distribution related areas has always been an important research focus. To get the distribution similarity between the words, the co-occurrence frequency, i.e. the frequency of two words appearing at the same time, has to be calculated. In traditional corpus environment, the contents of each document are analyzed to take the surrounding terms in a certain size window (the size is always set to 2) as co-occurrence terms. However, due to the feature of Web data as massiveness and high dynamicity, this method is difficult to be used in the Web environment. To address this problem, researchers consider to extend the word co-occurrence statistics window to the entire Web page. That is, the word co-occurrence is based on the entire Web page, and the word frequency is also based on the entire web page statistics ( a word can only be counted as one even if it appears several times in a page, while the traditional corpus counts the word frequency according to the actual appearance frequency in a document).

After enlarging the word co-occurrence statistics window to the entire web page, we can directly use the search engines to calculate vocabulary similarity. A query can include one term or two terms. When there is only one query word, the number of pages of the returned results reflects the number of pages which contain the query word in the entire Web corpus, i.e. the frequency of the query word. When there are two words, the number of the returned results reflects the co-occurrence frequency between the two query words. Web-based or corpus-based distribution similarity calculation is essentially the same. So many corpus-based similarity calculation methods can be applied to the former, such as the Jaccard coefficient, Dice coefficient, Pointwise Mutual Information (PMI), Google Similarity Distance (GND), etc. Their definition are shown as the formula （１）—（４）[18, 19]：

$$WebJaccard(q_1,q_2)=\begin{cases}0, & \text{if } hits(q_1\wedge q_2)\le c\\ \log_2\left(\dfrac{hits(q_1\wedge q_2)}{hits(q_1)\times hits(q_2)-hits(q_1\wedge q_2)}\right), & \text{otherwise}\end{cases} \tag{1}$$

$$WebDice(q_1,q_2)=\begin{cases}0, & \text{if } hits(q_1\wedge q_2)\le c\\ \log_2\left(\dfrac{2\times hits(q_1\wedge q_2)}{hits(q_1)\times hits(q_2)}\right), & \text{otherwise}\end{cases} \tag{2}$$

$$WebPMI(q_1,q_2)=\begin{cases}0, & \text{if } hits(q_1\wedge q_2)\le c\\ \log_2\left(\dfrac{hits(q_1\wedge q_2)/N}{\left(hits(q_1)/N\right)\times\left(hits(q_2)/N\right)}\right), & \text{otherwise}\end{cases} \tag{3}$$

$$NGD(q_1,q_2)=\frac{\max\left(\log\left(hits(q_1)\right),\log\left(hits(q_2)\right)\right)-\log\left(hits(q_1\wedge q_2)\right)}{\log N-\min\left(\log\left(hits(q_1)\right),\log\left(hits(q_2)\right)\right)} \tag{4}$$

Where $hits(q)$ is the number of the pages of the search results that contain the query word $q$, $hits(q_1\wedge q_2)$ is the number of the pages of the search results that contain the both query word $q_1$ and $q_2$, $N$ is the number of total pages the search engine have indexed (based on the Google's result, this is set to $10^{10}$ [19]), $c$ is a threshold used to filter out low frequency interference items ($c$ is set to 5 in this paper). The collection of the return page fragments (Snippet) also contains other useful information besides frequency. Note that formula (4) defines the similarity distance between $q_1$ and $q_2$, while this paper focuses on the similarity. So formula (4) is rewritten as:

$$NGS=1-NGD=\begin{cases}0, \text{ if } NGD>1\\ \dfrac{\log\left(N\times hits(q_1\wedge q_2)\right)-\log\left(hits(q_1)\times hits(q_2)\right)}{\log N-\min\{\log\left(hits(q_1)\right),\log\left(hits(q_2)\right)\}}\end{cases} \tag{5}$$

Study shows that the result is not satisfactory if directly using the above formula to calculate Web-based distribution similarity. This paper will use a two-way validation similarity strategy which is proposed by Chen etc, and on this basis improve the Co-occurrence Double-check Mode (CODC) [20].

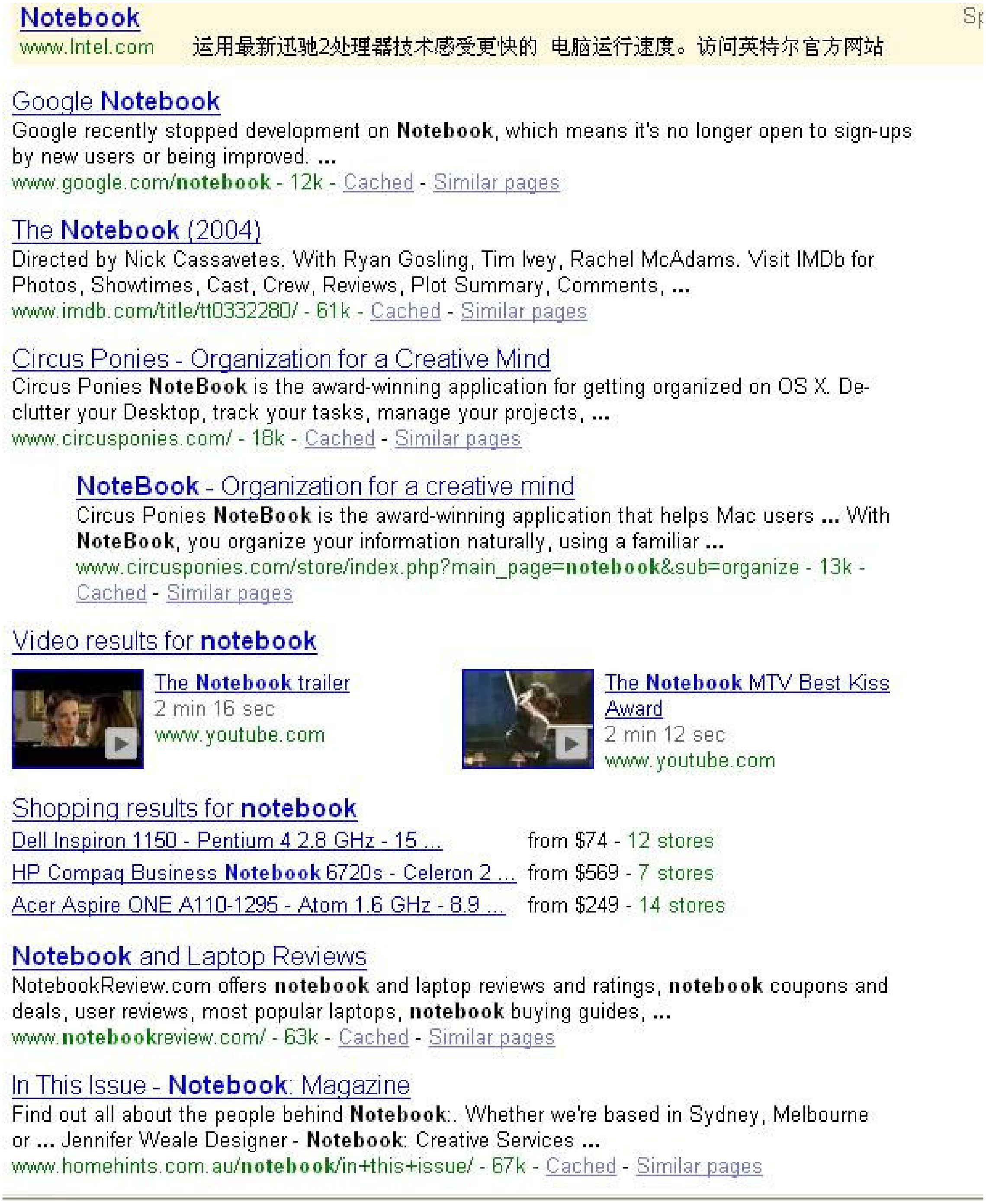
## 2.2 CODC Model

Co-occurrence Double-check Mode (below abbreviated as CODC Model) is a vocabulary similarity calculation model based on Web search engines: given two words x and y whose similarity needs to be calculated, adopting a search engine (Google search engine in this paper), first use query word X to obtain the search result D(X) which is the collection of web page fragments (hereinafter referred to as Snippet) containing the query word X; then use word Y as query word to get the search result $D(Y@X)$ in Snippet collection D(X), obviously $D(Y@X)\subseteq D(X)$. To do a similar operation with Y, we can get $D(Y)$, $D(X@Y)$, and $D(X@Y)\subseteq D(Y)$. Finally the similarity of X and Y can be calculated as

:

$$CODC(X,Y)=\begin{cases}0, & if\ \ |D(Y@X)|=0\forall|D(X@Y)|=0\\ e^{\lambda}:\lambda=\log\left(\dfrac{|D(Y@X)|}{|D(X)|}\times\dfrac{|D(X@Y)|}{|D(Y)|}\right)^{\partial}, & \text{Otherwise}\end{cases} \tag{6}$$

Where $|D(X)|$ and $|D(Y)|$ respectively corresponds to the number of web pages of search results of query word X or Y, $|D(Y @ X)|$ is the number of Snippets containing the query word Y in $D(X)$. $CODC(X,Y)$ is in the range [0,1], equal to the minimum value 0 while $|D(X @ Y)| = 0$ or $|D(Y @ X)| = 0$ and equal to the maximum value 1 while $|D(Y @ X)| = |D(X)|$ and $|D(X @ Y)| = |D(Y)|$.



Notebook
www.Intel.com 运用最新迅驰2处理器技术感受更快的 电脑运行速度。访问英特尔官方网站 Sp

Google Notebook
Google recently stopped development on **Notebook**, which means it's no longer open to sign-ups by new users or being improved. ...
www.google.com/**notebook** - 12k - Cached - Similar pages

The Notebook (2004)
Directed by Nick Cassavetes. With Ryan Gosling, Tim Ivey, Rachel McAdams. Visit IMDb for Photos, Showtimes, Cast, Crew, Reviews, Plot Summary, Comments, ...
www.imdb.com/title/tt0332280/ - 61k - Cached - Similar pages

Circus Ponies - Organization for a Creative Mind
Circus Ponies **NoteBook** is the award-winning application for getting organized on OS X. De-clutter your Desktop, track your tasks, manage your projects, ...
www.circusponies.com/ - 18k - Cached - Similar pages

NoteBook - Organization for a creative mind
Circus Ponies **NoteBook** is the award-winning application that helps Mac users ... With **NoteBook**, you organize your information naturally, using a familiar ...
www.circusponies.com/store/index.php?main_page=**notebook**&sub=organize - 13k - Cached - Similar pages

Video results for **notebook**

The **Notebook** trailer
2 min 16 sec
www.youtube.com

The **Notebook** MTV Best Kiss Award
2 min 12 sec
www.youtube.com

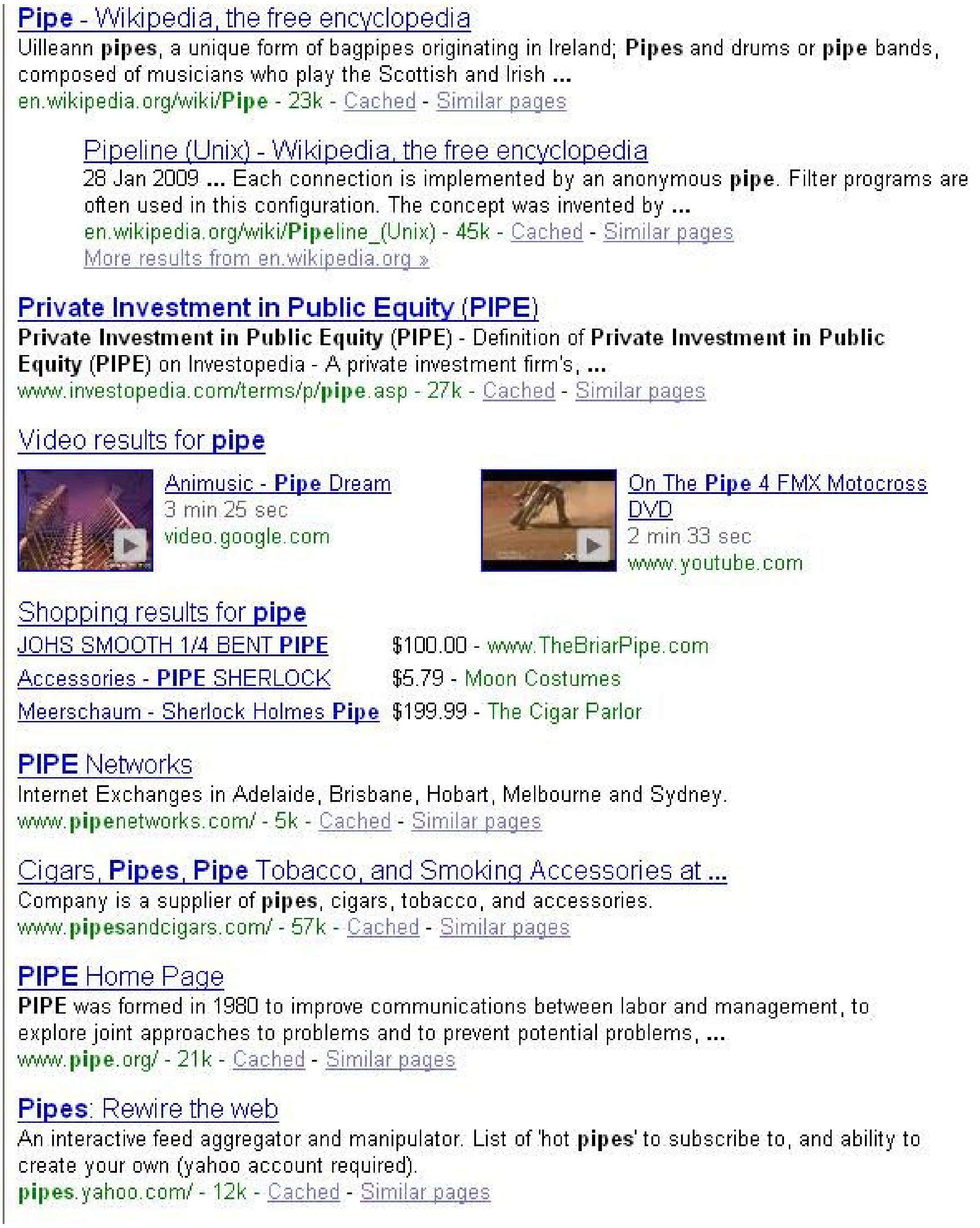
Shopping results for **notebook**

| | |
|---|---|
| Dell Inspiron 1150 - Pentium 4 2.8 GHz - 15 ... | from $74 - 12 stores |
| HP Compaq Business **Notebook** 6720s - Celeron 2 ... | from $569 - 7 stores |
| Acer Aspire ONE A110-1295 - Atom 1.6 GHz - 8.9 ... | from $249 - 14 stores |

Notebook and Laptop Reviews
NotebookReview.com offers **notebook** and laptop reviews and ratings, **notebook** coupons and deals, user reviews, most popular laptops, **notebook** buying guides, ...
www.**notebook**review.com/ - 63k - Cached - Similar pages

In This Issue - **Notebook**: Magazine
Find out all about the people behind **Notebook**:. Whether we're based in Sydney, Melbourne or ... Jennifer Weale Designer - **Notebook**: Creative Services ...
www.homehints.com.au/**notebook**/in+this+issue/ - 67k - Cached - Similar pages

Figure 2 The returned results of query word "Notebook" from Google search engine.



**Pipe** - Wikipedia, the free encyclopedia
Uilleann **pipes**, a unique form of bagpipes originating in Ireland; **Pipes** and drums or **pipe** bands, composed of musicians who play the Scottish and Irish **...**
en.wikipedia.org/wiki/**Pipe** - 23k - Cached - Similar pages

Pipeline (Unix) - Wikipedia, the free encyclopedia
28 Jan 2009 **...** Each connection is implemented by an anonymous **pipe**. Filter programs are often used in this configuration. The concept was invented by **...**
en.wikipedia.org/wiki/**Pipe**line_(Unix) - 45k - Cached - Similar pages
More results from en.wikipedia.org »

**Private Investment in Public Equity (PIPE)**
**Private Investment in Public Equity** (PIPE) - Definition of **Private Investment in Public Equity** (PIPE) on Investopedia - A private investment firm's, **...**
www.investopedia.com/terms/p/**pipe**.asp - 27k - Cached - Similar pages

Video results for **pipe**

Animusic - **Pipe** Dream
3 min 25 sec
video.google.com

On The **Pipe** 4 FMX Motocross DVD
2 min 33 sec
www.youtube.com

Shopping results for **pipe**

| | |
|---|---|
| JOHS SMOOTH 1/4 BENT **PIPE** | $100.00 - www.TheBriarPipe.com |
| Accessories - **PIPE** SHERLOCK | $5.79 - Moon Costumes |
| Meerschaum - Sherlock Holmes **Pipe** | $199.99 - The Cigar Parlor |

**PIPE** Networks
Internet Exchanges in Adelaide, Brisbane, Hobart, Melbourne and Sydney.
www.**pipe**networks.com/ - 5k - Cached - Similar pages

Cigars, **Pipes**, **Pipe** Tobacco, and Smoking Accessories at **...**
Company is a supplier of **pipes**, cigars, tobacco, and accessories.
www.**pipes**andcigars.com/ - 57k - Cached - Similar pages

**PIPE** Home Page
**PIPE** was formed in 1980 to improve communications between labor and management, to explore joint approaches to problems and to prevent potential problems, **...**
www.**pipe**.org/ - 21k - Cached - Similar pages

**Pipes**: Rewire the web
An interactive feed aggregator and manipulator. List of 'hot **pipes**' to subscribe to, and ability to create your own (yahoo account required).
**pipes**.yahoo.com/ - 12k - Cached - Similar pages

Figure 3 The returned results of query word "pipe" from Google search engine.

In the previous Web-based vocabulary similarity calculation approaches, both the occurrence frequency of query word and the co-occurrence frequency between the two query words are calculated overall based on the search results. Now, using double-check based on query results, CODC model can to some extent filter out those results differing greatly between the search results of the two query words, so that the vocabulary distribution similarity calculation and the true vocabulary distribution have a better goodness of fit. Although this method is simple and intuitive, there is a problem: the returned results reflect various kinds of semantic levels of the query word. For example, figure 2 shows the returned results of query word "Notebook" from Google search engine. Some Snippets describe the laptop computer "notebook", some are about movie “notebook”, and some are about book for writing notes, etc. Figure 3 the returned results fo query word “pipe” from Google search engine. Some Snippets describe the pipe as hollow cylinder, some Snippets describe it as a kind of music instrument, some Snippets describe it as a set of data processing elements connected in series, and some Snippets describe it as smoking accessory.

### 2.3 Revised CODC Model (RCODC)

In CODC model, the word semantic similarity measurement is directly based on the search results which may correspond to various kinds of semantic levels of the query word due to the diversity of the Web content. In addition, previous studies observed that when the X and Y have low similarity (but not 0), $D(X@Y)$ or $D(Y@X)$ often becomes 0, which leads to the final result of similarity being 0. While calculating the vocabulary similarity, simply setting the low vocabulary similarity to 0 leads to a certain degree of information loss. A revised model, RCODC, is proposed here to solve the above two problems. It improves the original CODC model from two aspects:

1. First classify the Snippet collections ( ( $D(X)$ and $D(Y)$ ) of the search results according to their themes, then get the information of occurrence frequency and the co-occurrence frequency based on the categories, and later calculate the vocabulary similarity with these information.
2. Adopt synonyms in WordNet to expand the double-check process. Set $D(X@Y)$ as the number of Snippets in $D(Y)$ of the query word Y or synonyms of Y (which can be retrieved from WordNet).

By the assumption that the search result $D(X)$ of the query word X can be divided into $n$ collections, $D_1(X)$ , $D_2(X)$ , … , $D_n(X)$ , corresponding to different themes $c_1, c_2, ..., c_n$, the formula CODC (6) can be rewritten as:

$$CODC^i(X,Y) = \begin{cases} 0, & if \ \ |D_i(Y@X)| = 0 \ \forall \ |D_i(X@Y)| = 0 \\ e^{\lambda} : \lambda = \log\left( \dfrac{|D_i(Y@X)|}{|D_i(X)|} \times \dfrac{|D_i(X@Y)|}{|D_i(Y)|} \right)^{\partial}, & \text{Otherwise} \end{cases} \quad (7)$$

Where, $D_i(X)$ is the Snippet collection of class $c_i$ in the search results of query word $X$ , and $\partial$ is a free adjustable parameter and set to 0.15 in this paper.

Model RCODC is defined as the similarity between word *X* and *Y* is:

$$RCODC(X,Y) = \arg\max_{\forall i}\left( CODC^i(X,Y) \right) \quad (8)$$

Therefore, for any given two words, $w_1$ and $w_2$, their similarity is:

$$sim(w_1, w_2) = RCODC(w_1, w_2) \quad (9)$$

## 3 Experiment and Result Analysis

### 3.1 Experimental data

Two standard datasets for word semantic similarity measurement testing are adopted in the experiment, one is R&G dataset[21 ] designed by Rubenstein and Goodenough, and the other is M&C dataset[22 ] built by Miller and Charles.

The R&G dataset was established by Rubenstein and Goodenough in the synonyms testing experiment in 1965. The R&G dataset contains 65 pairs of nouns, and 51 volunteers have been invited to score for each pair: evaluate the similarity of each pair of words with the score of 0.0 to 4.0, where the score of 4.0 means the ultimate similarity. And since the test mainly focuses on the normal language research instead of some specialized fields, these nouns are picked from commonly used English words, excluding any professional terms.

Miller and Charles repeated the Rubenstein and Goodenough's experiment later in 1991 with smaller data set. The data set which is named the M&C data set contains 30 pairs of words extracted from R&G dataset. These 30 word pairs are evenly picked from the score interval of [3,4], [1,3], and [0,1]. The above two testing data sets have been used for word semantic similarity measurement testing by a great number of researchers ever since they were established and have become the de facto standards[1, 2, 15, 23, 24] of this field. Thus, they are adopted in the following experiment.

### 3.2 Experimental Settings

Seven well-known search engines are used in the experiment, which are Google,Yahoo, MSN, AllTheWeb, AOL, Ask Jeeves, and Lycos. After inputting one word, the first 1000 records of the returned results from these 7 engines are respectively collected as the Statistical information source of this particular word, and then the word frequency information can be extracted from it. In addition, the snippets in the search results should be categorized before calculating word semantic similarity. Support Vector Machine(SVM) is selected in the experiment to categories snippets which applies the open source software packages Libsvm[25] provided by Chang and Lin.

In addition, the Snippet classification system experiment adopts the DMOZ Open Directory Project category architecture. There are 16 top-level directory structures: Arts, Business, Computers, Games, Health, Home, Kids and Teens, News, Recreation, Reference, Regional, Science, Shopping, Society, Sports and World.

Since the Reference, Regional and World categories have extensive definitions and the border of each category is relatively vague and not easy to judge, these three categories are excluded in the experiment.. That is to say only the other 13 categories are used which are noted as $C_i$ , $i = 1,...,13$. The training examples of each category are picked from its web pages. In the experiment, 5000 web pages are extracted from each category, and then the contexts that contain the category word are picked from each page as the tagging examples for the corresponding category (the context is three sentences long).

### 3.3 Experimental Results and Analysis

Word similarities calculated by different algorithms can not be compared directly, because their measurement methods are different. Thus Pearson product-moment correlation coefficient, ρ in short, is employed in this experiment as a testing standard to compare the quality of different algorithms, which is a popular testing method of word similarity calculation in the present. Moreover, to insure objectivity and impartiality, other word similarity calculation methods including LCH, LIN, WUP, JCN, LESK which are based on WordNet(2.1 edition) and supplied by Pedersen[26] software package are tested in the experiment besides the method proposed in this paper. However, due to the limited space only two of the best results: LCH and LIN are listed in the paper. In addition, WebJaccard, WebDice and WebPIM which are based on web engine (google) are tested, and only the best result WebPMI is listed. And Web search engines based GNS and CIDC methods also have been tested.

The results of the experiment are showed in Table 1 and Table 2. The results illustrate that the word similarity calculation method proposed in this paper has an even performance as the best result method LCH based on WordNet.

According to R&G data set, the value ρ calculated by RCODC is 82%, only 1% less than the highest value 83% achieved by LCH and much more than the others: 73%(LIN), 32%(PMI), 21% (GNS), 77%(CODC). According to M&G data set, similar results are obtained which shows that both RCOD and LCH get the highest 81%, more than the results of other calculation methods: 80%(LIN), 49%(PMI), 24%(GNS), 79%(CODC)

.

| | LCH | LIN | PMI | GNS | CODC | Ours |
|---|---|---|---|---|---|---|
| ρ | 0.83 | 0.73 | 0.32 | 0.21 | 0.77 | 0.82 |

Table 1. Experimental results according to R&G dataset

| | LCH | LIN | PMI | GNS | CODC | Ours |
|---|---|---|---|---|---|---|
| ρ | 0.81 | 0.80 | 0.49 | 0.24 | 0.79 | **0.81** |

Table 2. Experimental results according to M&G dataset

It can be seen from the above results that the RCODC model proposed in this paper performs as good as the methods based on WordNet, with one more advantage that the former can deal with the terms which are not covered in WordNet, such as "midday-noon" in the experiment, due to the massive data on the web. In addition, by comparing RCODC with CODC, it can be seen that CODC model simply sets the similarity value of low similarity word pairs as 0 while RCODC model solves this problem by expanding the check scope with using synonyms extracted from WordNet in double-check process.

## 4 Conclusions

To address the issues existing in the word semantic similarity measurement based on WordNet and Corpus, a Web search engines based word semantic similarity measurement model — RCODC is proposed in this paper. Three main features are as follows: First, the whole Web environment is taken as a corpus for word similarity calculation , and the statistical information can be obtained by search engines (Google in this paper). Thanks to the huge volumes of information in the Web, the low term coverage problem of WordNet and the sparse data problem of Corpus can be much improved. Second, after adopting the synonyms defined in WordNet, more information can be retrieved from the search results to support word similarity calculation. Third, the revised feature model is employed in Snippet category to improve the accuracy of Snippet category, and further improve the accuracy of word similarity calculation which is based on the Snippet category.

## Acknowledgement

The work has been partially supported by the Zhejiang Significant Science & Technology Project with Special and Priority Topic (No. 2007C13050), Zhejiang Natural Science Foundation of China (No. Y1080170, No. Y1090351) and Zhejiang Science & Technology Plan of China (No. 2009C31116).

*References:*

[1] Resnik P. *Using information content to evaluate semantic similarity in a taxonomy*. Proceedings of the 14th International Joint Conference on Artificial Intelligence, 1995, vol. 1 pp. 448-453.

[2] Jiang J. J.Conrath D. W. *Semantic similarity based on corpus statistics and lexical taxonomy*. Proceedings of International Conference on Research in Computational Linguistics, 1997, pp. 19-33.

[3] Lin D. *An information-theoretic definition of similarity*. Proceedings of the 15th International Conference on Machine Learning, 1998, pp. 296-304.

[4] Leacock ClaudiaChodorow Martin. *Combining local content and WordNet similarity for word sense identification*. WordNet: An Electronic Lexical Database, Chapter 11, 1998, pp. 265-283.

[5] Gurevych I. *Using the Structure of a Conceptual Network in Computing Semantic Relatedness*. Proceedings of the 2nd International Joint Conference on Natural Language Processing, 2005, pp. 767–778.

[6] Budanitsky A.Hirst G. *Evaluating wordnet-based measures of lexical semantic relatedness*. Computational Linguistics, 2006, vol. 32 pp. 13-47.

[7] Patwardhan S.Pedersen T. *Using WordNet-based Context Vectors to Estimate the Semantic Relatedness of Concepts*. space, 2006, vol. 1501,p. 1.

[8] Strube M.Ponzetto S. P. *WikiRelate! Computing semantic relatedness using Wikipedia*. Proc. of AAAI, 2006, vol. 6 pp. 1419–1424.

[9] Gabrilovich E.Markovitch S. *Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis*. Proceedings of the 20th International Joint Conference on Artificial Intelligence, 2007, pp. 6–12.

[10] Hughes T.Ramage D. *Lexical semantic relatedness with random graph walks*. Proceedings of EMNLP, 2007, vol. 7

[11] Fellbaum C. *Wordnet: An Electronic Lexical Database*: MIT Press, 1998.

[12] Lee J. H. *Information Retrieval Based on Conceptual Distance in Is-A Hierarchies*. Journal of Documentation, 1993, vol. 49 pp. 188-207.

[13] Wu Z.Palmer M. *Verb semantics and lexical selection*. Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, 1994, pp. 133–138.

[14] Sussna M. *Word sense disambiguation for free-text indexing using a massive semantic network*. Proceedings of the second international conference on Information and knowledge management, 1993, pp. 67-74.

[15] Lin D. *An information-theoretic definition of similarity*. Proceedings of the 15th International Conference on Machine Learning, 1998, pp. 296–304.

[16] Weeds J. E.University of Sussex, *Measures and Applications of Lexical Distributional Similarity*:

University of Sussex, 2003.

[17] Evert S., *The Statistics of Word Cooccurrences Word Pairs and Collocations*: s. n, 2005.

[18] Bollegala D., Matsuo Y.Ishizuka M., *WebSim: A Web-based Semantic Similarity Measure*, 2007.

[19] Rudi L. CilibrasiPaul M. B. Vitanyi. *The google similarity distance*. Transactions on Knowledge and Data Engineering, 2007, vol. 19 pp. 370-383.

[20] Chen H., Lin M.Wei Y. *Novel association measures using web search with double checking*. Proc. of the COLING/ACL 2006, 2006, pp. 1009–1016.

[21] Rubenstein H.Goodenough J. B. *Contextual correlates of synonymy*. Communications of the ACM, 1965, vol. 8 pp. 627-633.

[22] Miller G. A.Charles W. G. *Contextual correlates of semantic similarity*. Language and Cognitive Processes, 1991, vol. 6 pp. 1-28.

[23] Budanitsky A.Hirst G. *Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures*. Workshop on WordNet and Other Lexical Resources, Second meeting of the North American Chapter of the Association for Computational Linguistics, 2001, pp. 29–34.

[24] Jarmasz M.Szpakowicz S. *Roget's Thesaurus and Semantic Similarity*. Recent Advances in Natural Language Processing III Selected Papers from RANLP 2003, 2004,

[25] Chang C. C.Lin C. J. *LIBSVM: a library for support vector machines*. Software available at http://www.csie.ntu.edu.tw/cjlin/libsvm, 2001, vol. 80 pp. 604-611.

[26] Pedersen T., Patwardhan S.Michelizzi J. *WordNet:: Similarity-Measuring the Relatedness of Concepts*. Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI-04), 2004, vol. 428