Semantic Cosine Similarity

Faisal Rahutomo*, Teruaki Kitasuka, and Masayoshi Aritsugi
Graduate School of Science and Technology, Kumamoto University

Abstract- Cosine similarity is a widely implemented metric in information retrieval and related studies. This metric models a text as a vector of terms and the similarity between two texts is derived from cosine value between two texts' term vectors. Cosine similarity however still can't handle the semantic meaning of the text perfectly. This paper proposes an enhancement of cosine similarity measurement by incorporating semantic checking between dimensions of two term vectors. This strategy aims to increase the similarity value between two term vectors which contain semantic relation between their dimensions with different syntax. Experimental result shows our proposal yields a promising result.

Index Terms- cosine similarity, semantic, WordNet.

I. Introduction

COSINE similarity is a widely implemented metric in information retrieval and related studies. This metric models a text document as a vector of terms. By this model, the similarity between two documents can be derived by calculating cosine value between two documents' term vectors [5]. Implementation of this metric can be applied to any two texts (sentence, paragraph, or whole document). In search engine case, similarity value between user query and documents are sorted from the highest one to the lowest one. The higher similarity score between document's term vector and query's term vector means more relevancy between document and query.

Cosine similarity for similarity measurement between document and user query should accommodate to the word's meaning. Cosine similarity however still can't handle semantic meaning of the text perfectly. The implementation of cosine similarity measurement between two term vectors syntactically sometimes yields unreliable result. Syntax matching may not be able to meet the difference of semantic meaning problem. For further process, i.e., information retrieval system, it may produce false result and cause degrading in its performance.

Studies on semantic measurement or semantic similarity between words have been done. The most common method utilizes a lexical database as a semantic network, i.e., WordNet. The similarity between two concepts can be derived based on WordNet's exploration [2][7].

This paper proposes a simple enhancement of cosine similarity by including semantic checking between dimensions of two term vectors. This checking uses one of semantic word similarity methods such as Wu and Palmer [6] which explores WordNet's semantic network. This strategy aims to increase the similarity value between two term vectors. It contains semantic relation between their dimensions against different syntax. Furthermore, this research aims to make the result value be more reasonable than typical cosine similarity measurement, based on human judgment.

II. Related Work

A. Cosine Similarity

In document-query cases, a document can be represented as a term vector that the vector's dimensions refer to the terms available in the document [5]. Dimension's value is occurrence of term inside a document. A document can be described as a vector form as:

$$\vec{d} = (w_{d0}, w_{d1}, \dots, w_{dk}) \quad (1)$$

As same as the document, the query of term can be described as a vector form as:

$$\vec{q} = (w_{q0}, w_{q1}, \dots, w_{qk}) \quad (2)$$

where w_{di} and w_{qi} $(0 \le i \le k)$ are float numbers indicating the frequency of each term inside a document, while each vector's dimension corresponds to a term available in the document.

Based on vector similarity, similarity between two vectors can be defined as:

$$Sim(\vec{q}, \vec{d}) = \frac{\vec{q}.\vec{d}}{|\vec{q}||\vec{d}|} = \frac{\sum_{k=1}^{t} w_{qk} \times w_{dk}}{\sqrt{\sum_{k=1}^{t} (w_{qk})^2} \cdot \sqrt{\sum_{k=1}^{t} (w_{dk})^2}} \quad (3)$$

B. WordNet

WordNet aims to model lexical knowledge of English native speaker. Smallest unit in WordNet is logical group named synset, which represent specific meaning of a word (sense). Synsets have explicit semantic relations each other [4].

C. Semantic Similarity between Words

The similarity meaning between two words can be measured by involving knowledge based lexical semantic like WordNet. It can be measured using taxonomy based semantic word similarity. The methods are: Wu and Palmer, Lin, Resnick, Jiang and Conrath, and Lesk [2][7]. We can involve one of those methods based on some considerations such as path inside taxonomy, density, or combination of them [7].

As example, Wu and Palmer similarity [6] is one of the most popular methods due to its computational speed [3]. Wu and Palmer similarity measurement is a similarity measurement between two nodes in taxonomy based on edge counting as mentioned in Fig.1 and (4). Shorter edge-distance between two edges gives more similarity value. Based on Fig.1, Wu and Palmer similarity formulation between nodes C1 and C2 is written as:

$$Sim(C1, C2) = \frac{2 \times N_3}{N_1 + N_2 + (2 \times N_3)} \quad (4)$$

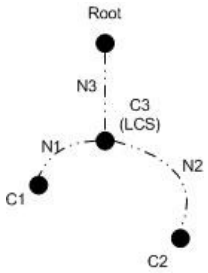
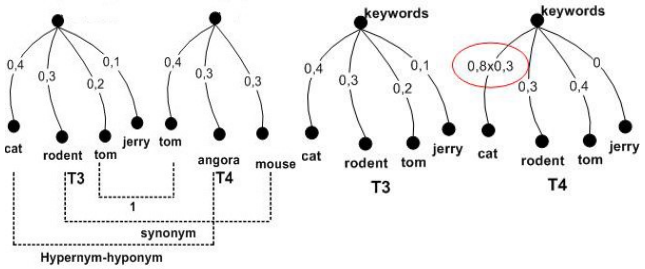


Figure 1.Wu and Palmer similarity concept.



(a) Case study **(b)** Enhanced situation

Figure 2. Semantic cosine similarity.

III. Semantic Cosine Similarity

A. Identified Problem

The cosine similarity measurement has a disadvantage. For instance, a case study is drawn (using weighted tree form [1]) in Fig.2a. There are two term vectors with some semantic relation on their dimensions, i.e., synonym relation and hypernym-hyponym relation. If we implement cosine similarity to this case, we will get a low similarity result.

First step, both of term vectors equalize their dimension. Cosine similarity of two vectors can be applied for vectors with same dimension only. For the above case, a new dimension is built with zero value in a term vector if there is a dimension with no pair in another term vector. $\vec{q}$ =(0.4,0.3,0.2,0.1,0,0) and $\vec{d}$ =(0,0,0.4,0,0.3,0.3) after dimension equalization. Cosine similarity measurement between the query vector and the document vector is:

$$Sim(\vec{q},\vec{d}) = \frac{(0.4\times 0)+(0.3\times 0)+(0.2\times 0.4)+(0.1\times 0)+(0\times 0.3)+(0\times 0.3)}{\sqrt{0.4^2+0.3^2+0.2^2+0.1^2}.\sqrt{0.4^2+0.3^2+0.3^2}}$$

$$Sim(\vec{q},\vec{d}) = 0.250$$

The similarity value is too small. It is different from our human judgment measurement. We know "rodent" is synonym of "mouse" and "angora" is-a-kind-of "cat", as drawn in Fig.2a. It is not fair giving similarity value 0.250 if we consider the semantic relation between their dimensions.

B. Enhancement

Based on the identified problem, this research proposes the following steps:

1. Dimension equalization of both term vectors.
2. If there is any synonym pair with different syntax between two term vectors, choose the first one as dimension name for both vectors. In Fig.2b, rodent is chosen as dimension's name for "rodent-mouse" synonym pair.
3. If there is a dimension in first term vector with hypernym-hyponym relation toward second term vector, choose one as dimension name. Another dimension will not be chosen as dimension name. In Fig.2b, cat is chosen as dimension's name for "cat-angora" hypernym-hyponym relation.
4. Recalibrate dimension value of dimensions with hypernym-hyponym relation by this formula:

$$n = p \times s \quad (5)$$

where n is new dimension value, p is previous similarity value, and s is hypernym-hyponym similarity value. We can derive this value by WordNet exploration and calculate the semantic similarity between words by a method such as described in Section II.C. In Fig.2b, the hypernym-hyponym similarity assumption value of "cat-angora" is 0.8.

C. Result

Through enhancement steps for case problem in Fig.2a, $\vec{q}$=(0.4,0.3,0.2,0.1), $\vec{d}$ =(0.24,0.3,0.4,0), and cosine similarity measurement between query vector and document vector is:

$$Sim(\vec{q},\vec{d}) = \frac{(0.4\times 0.24)+(0.3\times 0.3)+(0.2\times 0.4)+(0.1\times 0)}{\sqrt{0.4^2+0.3^2+0.2^2+0.1^2}.\sqrt{0.24^2+0.3^2+0.4^2}}$$

$$Sim(\vec{q},\vec{d}) = 0.875$$

We get similarity score 0.875 for case problem in Fig.2a, through the enhancement scenario in Section III.b, as shown in Fig.2b. The result is more reasonable than typical cosine similarity if we consider the semantic relation among vector's dimensions.

IV. Conclusion

This paper proposed an enhancement of cosine similarity between two term vectors. This enhancement considers not only standard vector operation but also the semantic relation between vector's dimensions. Our experiment yielded more reasonable value, based on human judgment.

V. References

[1] V.C. Bhavsar, H. Boley, L. Yang, "A Weighted-Tree Similarity Algorithm for Multi-agent System in E-Business Environments," *Computational Intelligence*, vol. 20, no. 4, 2004, pp. 584–602.

[2] C. Corley and R. Mihalcea, "Measuring the Semantic Similarity of Text," in *Proceedings of the ACL workshop on Empirical Modeling of Semantic Equivalence and Entailment*, 2005, pp. 13–18.

[3] A. Madylova and S.G. Oguducu, "A taxonomy based semantic similarity of documents using the cosine measure," in *Proceeding of International Symposium on Computer and Information Sciences*, 2009, pp. 129–134.

[4] G.A. Miller, R. Beckwith, C. Fellbaum, D. Gross and K. Miller, "WordNet: An on-line lexical database," *International Journal of Lexicography*, vol. 3, 1990, pp. 235–244.

[5] G. Salton and C. Buckley, "Term-weighting Approaches in Automatic Text Retrieval," *Information Processing and Managemen*t, vol.24, no.5, 1988, pp.513–523.

[6] Z. Wu and M. Palmer, "Verb Semantic and Lexical Selection," in Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistic, 1994, pp. 133–138.

[7] L. Yang, V.C. Bhavsar, H. Boley, "On Semantic Concept Similarity Methods," in *Proceedings of International Conference on Information & Communication Technology and System*, 2008, pp. 4–11.