Application of Topic Based Vector Space Model with WordNet

Adi Wibowo
Informatics Department
Petra Christian University
Surabaya, Indonesia
adiw@peter.petra.ac.id

Andreas Handojo
Electrical Department
Petra Christian University
Surabaya, Indonesia
handojo@peter.petra.ac.id

Albert Halim
Informatics Department
Petra Christian University
Surabaya, Indonesia

***Abstract*—Topic Based Vector Space Model (TVSM) proposed a new vector space that its dimensions is composed of topics. Every term and document is represented by vectors inside this vector space. By using topics as dimensions TVSM tries to overcome word-mismatch between terms with similar topics in finding relevant documents to query. This study proposes to develop relations between terms using WordNet and thesaurus to help TVSM calculating similarity between documents. Relations between terms are represented by relation score. This study proposes a way to find optimal relation score for a set of documents. To help indexing documents with multi language terms this study also proposes to use dictonary to expand query terms.**

Topic based vector space model, wordnet, dictionary

I. INTRODUCTION

Search engine is designed to find documents relevant to user queries. One of a method used in measuring similarity is Vector Space Model (VSM). VSM assumes every unique term from documents can be represented by each dimension of vector [1]. This method also has the assumption that each term in the document are independent. A term "lion" has no relationship or similarity with "animal", or "car" with "automotive". This condition is called word mismatch. Word mismatch is one of factor that causes a decrease in recall and lead to less accurate document similarity calculation process.

To overcome the problem Topic Based Vector Space Model (TVSM) has the assumption that each term has a relationship with other terms. Their proximity can be measured by using relationship from thesaurus [2]. Term generalization can also use dictionary, so it is possible to search multi language document and still provide relevant documents to user queries.

This paper tries to propose the application of WordNet as thesaurus and english-indonesian dictionary to provide terms and relationships needed by TVSM.

II. VECTOR SPACE MODEL AND TVSM

In Vector Space Model every document is represented by a vector in vector space. Each vector dimension represents every unique term found in all documents. Similarity between documents is calculated by measuring angle between their vector. In Vector Space Model documents and query are represented as equation 1.

978-1-4244-9983-0/11/$26.00 ©2011 IEEE

$$d_i = (w_{1,i}, w_{2,i}, w_{3,i}, \ldots, w_{t,i})$$
$$q = (w_{1,q}, w_{2,q}, w_{3,q}, \ldots, w_{t,q}) \quad (1)$$

Similarity measurement between document and query is performed by using cosine similarity measure.

$$Sim(q, di) = \frac{\sum_i w_{q,j} w_{i,j}}{\sqrt{\sum w_{q.j}^2}\sqrt{\sum w_{i.j}^2}} \quad (2)$$

$w_{q,j}$ is a weight of term j from query q, and $w_{i,j}$ is a weight of term j from document i. Each weight is calculated from tf*idf value.

$$w_{i,j} = tf_{i,j} * \log\left(\frac{D}{df_j}\right) \quad (3)$$

$tf_{i,j}$ is a number of term j in a document i, d_{fj} is a number of documents that contain term j, and D is total number of documents.

Topic Based Vector Space Model assumes another form of vector. If the vector dimension of VSM represents terms, the vector dimension of TVSM represents fundamental topic of terms [2]. Terms in TVSM is also represented by the vector. There is *d* dimensional space *R* which only has positive axis-intercepts.

$$R \epsilon \boldsymbol{R}_{\geq 0}^{d} \; with \; d \; \epsilon \; \boldsymbol{N}_{>0} \quad (4)$$

Each term $i \in \{1, \ldots, n\}$ is represented by *term vector* $\vec{t_i}$ and has term weight between 0 and 1. Term weight in TVSM is defined as:

$$\vec{t_i} = (t_{i1}, t_{i2}, \ldots, t_{id}) \; \epsilon \; R \quad (5)$$

$$\vec{t_i}| = \sqrt{t_{i1}^2 + t_{i2}^2 + \ldots + t_{id}^2} \; \epsilon \; [0; 1] \quad (6)$$

And term correlation is defined as :

$$\overrightarrow{t_i}\overrightarrow{t_j} = |\overrightarrow{t_i}| \cdot |\overrightarrow{t_j}| \cdot \cos(\omega_{ij}) \quad (7)$$

$$i, j \in \{1, \dots, n\}$$

$$i : |\overrightarrow{t_i}| \in [0; 1]$$

$$\omega_{ij} \in [0^o; 90^o]$$

A document $k \in \{1, \dots, m\}$ is represented by vector $\overrightarrow{d_k} \in R$, where vector length is standardized to one.

$$\overrightarrow{d_k} = \frac{1}{\vec{\delta}_k}\vec{\delta}_k \quad \rightarrow \quad |\vec{d}_k| = 1$$

$$\vec{\delta}_k = \sum_{i=1}^{n} e_{ki}\overrightarrow{t_i} \quad (8)$$

while

e_{ki} = number of term *i* in document *k*

With this document representation, similarity between document k and l is defined as scalar product from vector documents.

$$\overrightarrow{d_k}\overrightarrow{d_l} = |\overrightarrow{d_k}| \cdot |\overrightarrow{d_l}|. \text{Cos}(\omega_{kl}) = \cos(\omega_{kl})$$

$$\text{where } \overrightarrow{d_k} = \overrightarrow{d_l} = 1$$

$$\overrightarrow{d_k}\overrightarrow{d_l} = \frac{1}{|\vec{\delta}_k|}\vec{\delta}_k \frac{1}{|\vec{\delta}_l|}\vec{\delta}_l = \frac{1}{|\vec{\delta}_k||\vec{\delta}_l|}\sum_{i=1}^{n}\sum_{j=1}^{n} e_{ki}e_{lj}\vec{t}_i\,\vec{t}_j \quad (9)$$

The length of document vector $\vec{\delta}_k$ is defined as:

$$|\vec{\delta}_k| = \left|\sum_{i=1}^{n} e_{ki}\overrightarrow{t_i}\right| = \sqrt{\left(\sum_{i=1}^{n} e_{ki}\overrightarrow{t_i}\right)^2}$$

$$= \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} e_{ki}e_{kj}\vec{t}_i\,\vec{t}_j} \quad (10)$$

III. WordNet

WordNet is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. WordNet provides a network of meaningfully related words and concepts [3].

Synsets are term lists or collocation that have the same meaning and their use in particular context are interchangeable. Relations between terms used to describe relation between synsets are:

1. Hypernym. A word whose meaning denotes a superordinate or superclass. Animal is a hypernym of dog.
2. Hyponym. A word whose meaning is included in that of another word. Hyponym is an opposite of hypernym.
3. Holonym. A word for the whole of which other words are part. Building is a holonym of window.
4. Meronym. A word that names a part of a larger whole. Window is a meronym of building.
5. Coordinate term. A word is a coordinate term of other word if both words share similar hypernym. Wolf is a coordinate term of dog, and dog is a coordinate term of wolf.

Example of WordNet synsets are shown at table 1.

TABLE I. WordNet Synset Examples

Terms Related to Marketing	
Type of Relation	***Related Words***
Synonym	commerce, commercialism, mercantilism
Hypernym	commerce, commercialism, mercantilism, shopping
Hyponym	bait and switch, private treaty, dumping, distribution channel, channel
Holonym	marketing
Meronym	promotion, publicity, promotional material, packaging,
Coordinate term	commerce, commercialism, mercantilism, shopping

IV. Implementation

This study proposed to provide terms relations to TVSM by using WordNet from Princenton University, and Indonesian Thesaurus from Indonesian National Education Department Language Center. To create relationship between the two languages, English and Indonesian, this study used Stardict Dictionary.

A. Preparation

Every term found from documents will be checked to eliminate stop words. Stop words is a list of common or general terms (e.g., prepositions, and articles) that are not significant because they appear in too many records. Examples of stop words in English are 'a', 'the', 'an', 'for', 'of', etc. Example of stop words in Indonesian Language are 'di', 'ke', 'dari', 'bahwa', 'pada', etc. To remove stop words this study used English stop word list from Gerard Salton and Chris Buckley [4]. As for Indonesian Language, this study used a list from Indonesian Grammar from Moeliono [5].

The resulting terms will be stemmed using Porter Stemmer for English terms [6], or Indonesian stemming algorithm from Nasief and Adriani for Indonesian terms. For English terms before processed by stemmer, they will be checked if they are irregular verbs. Irregular verbs have to be restored to their base form.

B. Developing Relations Between Terms

WordNet synsets and Indonesian Thesaurus are used to create relations between terms. A relation is a score that represents proximity between terms. An optimum score will be determined after several testing. To develop relations between terms, several condition applied:

- Stardict is used to create relationship between English terms and Indonesian terms.

- The score of relation is assumed to be between 0 and 1.
- A term is always a word.
- If a term consists of one word is related to other terms consists of multiple words then the score of relation is divided evenly between each word.

When a query is entered into system several steps are performed:

- If a query is consists of more than one word then the query is separated into multiple terms.
- Every query term will go through preparation steps first.
- By using Stardict similar words from other language will be retrieved. This set of words will be called expanded terms. The query now consists of expanded terms.
- Related terms to each term from a query is determined by using synsets from WordNet and Indonesian Thesaurus. The relation score of each interrelated terms is already predefined using tests at chapter IV.C. For every other terms that has no relationship with terms from query will have relation score equal to zero.
- By using related terms and relation score TVSM is performed to find relevant documents to a query.

C. *Finding Optimum Relation Score*

In TVSM, e_{ki} is defined as the number of term i in document k. e_{ki} is used to calculate the weight of document k and similarity between documents. This study proposed to use of weighted e_{ki} by multipying e_{ki} with inverse document frequency (IDF). IDF measures the importance of a term in a document.

$$idf_t = log \frac{|D|}{|\{d \in D | t \in d\}|} \quad (11)$$

where

idf_t : *IDF* for *term* t
$|D|$: the total number of documents
$|\{d \in D | t \in d\}|$: number of documents where the term t appears

The method used in these tests is by searching a set of documents and measures Rank Weighted Average (RWA) scores. The assumption used by RWA is the more relevant the document it should be placed at the top rank. The lower score of RWA will provide better search engine results.

$$RWA = average\left(\frac{document\ rank\ in\ search\ result}{document\ rank\ in\ relevant\ assessment}\right) \quad (12)$$

The tests conducted to 350 documents with nine different relation score candidates between 0.1 and 0.9 with 0.1 interval to find which number provide better results. Each relation score is represented by R1 to R9 respectively. The queries used by the test is shown at table 2.

TABLE II. QUERIES USED BY TESTS

ID Query	Query terms
Q1	memilih mobil idaman
Q2	merapi mountain eruption
Q3	desain arsitektur rumah
Q4	supply chain management
Q5	data minings

The result of each test is shown at figure 1 to figure 5.

Figure 1. RWA from testing using query 1.

Figure 2. RWA from testing using query 2.

Figure 3. RWA from testing using query 3.

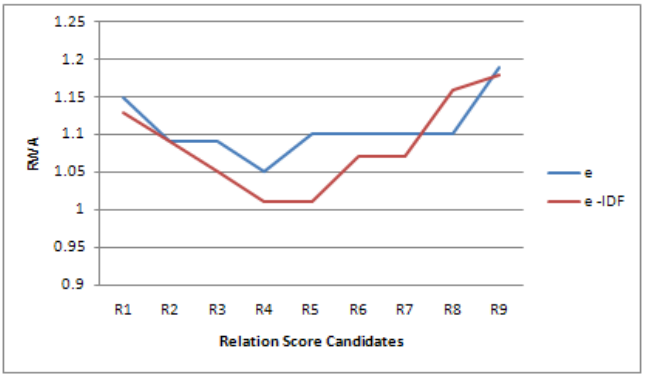

Figure 4. RWA from testing using query 4.

Figure 5. RWA from testing using query 5.

The average score of RWA from five tests shown at table III and figure 6 shows that e-IDF can help TVSM to obtain better RWA scores . These tests also shows that relation score R1 (0.1) provides better RWA compared with other relation scores.

	e	***e-IDF***
R1	1.929623	1.7516
R2	2.079487	1.944
R3	2.246713	2.1768
R4	2.510113	2.3986
R5	2.657908	2.5666
R6	2.881082	2.8124
R7	2.996781	2.97
R8	2.99821	3.2138
R9	3.280442	3.2702
MEAN	**2.62004**	**2.567111**

TABLE III. AVERAGE SCORES OF RWA FROM FIVE TEST.

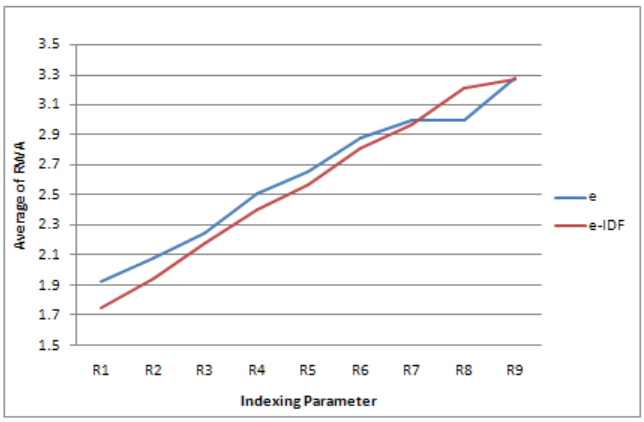

Figure 6. The average of RWA from five test for each relation score candidate.

V. CONCLUSION

This study proposes the use of WordNet and Indonesian thesaurus to build relations between terms and also propose a way to provide relation scores between terms. This study also proposes the use of dictionary to bridge relations between terms from different languages.

REFERENCES

[1] G. Salton, Automatic Information Organization and Retrieval, 1968

[2] J. Becker, D. Kuropka. "Topic based vector space model." USA: Business Information System, Proceesings of BIS, Colorado Springs. 2003

[3] Princeton University "About WordNet." WordNet. Princeton University. 2010. http://wordnet.princeton.edu

[4] C. Bukley dan G. Salton, Stopword List, Cornell University

[5] A.M.Moeliono, et.al. "Indonesian Grammar". Balai Pustaka: Department of Education and Cultures, 1988.

[6] M. Porter, "The Porter Stemming Algorithm". Retrieved May 24, 2011 from http://snowball.tartarus.org/algorithms/porter/stemmer.html.