Word Semantic Similarity based on document's title

Mohamed Said Hamani
Department of STIC
University of M'sila
M'sila, Algeria
saidhamani@hotmail.com

Ramdane Maamri
Lire Laboratory
University of Constantine 2
Constantine, Algeria
rmaamri@yahoo.fr

Abstract— **Measuring similarity between words using a search engine based on page counts alone is a challenging task. Search engines consider a document as a bag of words, ignoring the position of words in a document. In order to measure semantic similarity between two given words, this paper proposes a transformation function for web measures along with a new approach that exploits the document's title attribute and uses page counts alone returned by Web search engines. Experimental results on benchmark datasets show that the proposed approach outperforms snippets alone methods, achieving a correlation coefficient up to 71%.**

Keywords— Semantic Similarity; Decision Making; Title; Page Count; Text Mining;

I. INTRODUCTION

Semantic similarity between words is essential for many tasks and has been employed in various areas such as, Information Retrieval, Artificial Intelligence, Linguistics and Knowledge Engineering. Measuring the semantic similarity or dissimilarity (distance) between words is a process of quantifying the relatedness between the words using information sources [1].

Word similarity is a special case or a subset of word relatedness [2]. Based on information sources existing work on determining word relatedness is broadly categorized into three major groups [2]: corpus-based, knowledge-based and hybrid methods.

Semantic similarity between words changes over time and across domains. New words are constantly being created as well as new senses are assigned to existing words [3]. Manually maintaining thesauri to capture these new words and senses is costly if not impossible [3].

Regarding the Web as a live corpus has become an active research topic recently. Simple, unsupervised models demonstrably perform better when n-gram counts are obtained from the Web rather than from a large corpus [4]. Web search engines provide an efficient interface to access its massive store of information and return page counts and snippets for a given query.

Page count of a query is an estimate of the number of pages that contain the query words returned from a search engine. Snippets, a brief window of text extracted by a search engine around the query term in a document, provide useful information related to the local context of the query term [4].

The web search engine based approaches for measuring semantic between words can be categorized to page counts based approaches [9, 11], snippets based approaches [12, 13] and hybrid approaches [3, 4].

Page counts based approaches use the page counts alone returned from search engine as co-occurrence statistics to compute the semantic similarity between words.

One of the drawbacks of using page counts alone as a measure of co-occurrence of two words in a document is that page counts alone methods ignore word positions in a page considering the document as a bag of words.

Snippets based approaches use snippets, a brief window of text extracted by a search engine around the query term in a document.

Hybrid web search engine based approaches use the page count and the snippets in order to measure semantic similarity.

In order to measure semantic similarity between two given words, the latter should be considered in a specific context. Normally, co-occurrence is considered in a specific context or in a window of limited sizes such 3 to 7 words before or after a target word [7]. The strongest co-occurrence is the bi-gram in a short text.

Our idea is to find an attribute that is good enough to describe the content of a document and short enough for the co-occurrence to be considered. One of the attributes that is short and identifies the content of a document is the document's title.

In this paper, we propose a transformation function for web measures along with a novel approach based on page counts alone using the title of a document to measure semantic similarity between two given words. The paper study the URL attribute as well.

II. PAGE COUNT BASED MEASURES OF WORD RELATEDNESS

There is a relatively large number of co-occurrence measures in the literature such as Jaccard [4, 8], Overlap

(Simpson) [4], Dice [8], PMI (Point-wise mutual information) [4, 9] and Normalized Google Distance "NGD" [3, 4]. We use the notation count (t_1), count (t_2) as page counts returned by the search engine for the term "t_1", "t_2" respectively; count (t_1,t_2) represents the page counts for the query with terms "t_1" AND "t_2".

A. Jaccard Coefficient

The Jaccard coefficient [3, 8] measures similarity between two sets, and is defined as the size of the intersection divided by the size of the union of the two sets:

$$Jaccard(t_1,t_2) = \frac{count(t_1,t_2)}{count(t_1)+count(t_2)-count(t_1,t_2)} \qquad (1)$$

The Jaccard coefficient represents the maximum likelihood estimate of the ratio of the probability of finding a web document where terms t_1 and t_2 co-occur over the probability of finding a web document where either t_1 or t_2 occurs [2].

B. Dice coefficient.

Dice coefficient [2, 3] is defined as:

$$Dice(t_1,t_2) = \frac{2*count(t_1,t_2)}{count(t_1)+count(t_2)} \qquad (2)$$

The Dice coefficient represents the maximum likelihood estimate of the ratio of twice the probability of finding a web document where terms t_1 and t_2 co-occur over the probability of finding a web document where either t_1 or t_2 or both occurs [2].

C. Simpson Coefficient

Simpson or overlap coefficient [2, 3] is defined as:

$$Simpson(t_1,t_2) = \frac{count(t_1,t_2)}{Min(count(t_1),count(t_2))} \qquad (3)$$

The Simpson coefficient represents the maximum likelihood estimate of the ratio of the probability of finding a web document where terms t_1 and t_2 co-occur over the probability of finding a web document where the term with the lower frequency occurs [2].

D. Cosine Coefficient

Cosine coefficient is defined as:

$$Cosine(t_1,t_2) = \frac{count(t_1,t_2)}{\sqrt{count(t_1)*count(t_2)}} \qquad (4)$$

The Cosine coefficient compares the probability of observing the terms t_1 and t_2 together to the square root probabilities of observing t_1 and t_2 independently.

E. Pointwise Mutual Information

Pointwise Mutual Information (PMI) [3, 9] is defined as:

$$PMI(t_1,t_2) = \log_2\left(\frac{count(t_1,t_2)/N}{(count(t_1)/N)*(count(t_2)/N)}\right) \qquad (5)$$

PMI between two terms t_1 and t_2 compares the probability of observing the two terms together to the probabilities of observing t_1 and t_2 independently [2].

F. Normalized Google Distance (NGD)

The Normalized Google Distance [2, 11] is defined as:

$$NGD(t_1,t_2) = \frac{Max(\log(count(t_1)),\log(count(t_2)))-\log(counts(t_1,t_2))}{\log(N)-Min(count(t_1),count(t_2))} \qquad (6)$$

N is the number of documents indexed by the search engine.

NGD is a distance where the values of Equation (6) are unbounded, ranging from 0 to ∞. In order to transform NGD to a similarity measure, with values between 0 and 1, [14] defined NGD' similarity as:

$$NGD'(t_1,t_2) = e^{-2*NGD(t_1,t_2)} \qquad (7)$$

III. Transformation function for web similarity measures

Web search engines provide page counts of a query as an estimate of the number of pages that contain the query words. New documents are constantly being created on the web and indexed by the search engine. In a very dynamic environment, as the web, the propagation of uncertainty of page counts on the uncertainty of functions based on them conduct to undesirable results for formulas (1), (2),(3) and (4).

In this section we introduce a general transformation function to similarity measures to compute semantic relatedness using web page counts.

We define the general transformation function ξ as:

$\xi: R^+ \rightarrow R^+$ where

$$\forall x \in R^+ \xrightarrow{\xi} \xi(x) = \begin{cases} q^{n\log_p(x)} & x>0 \\ 0 & x=0 \end{cases} \qquad (8)$$

p, q: real numbers $p\neq 0$, $p\neq 1$ $q\neq 0$ and $q\neq 1$.

n: a real number $n\neq 0$ ($n\in R^*$).

As an application to our function (ξ), the transformation defined in [14] on NGD distance can be seen as a simple substitution into our transformation ξ with q=e (e: is the Napier's constant), n=-2 and p=10.

PMI defined in formulas (5) can be defined as the $\log_q$ transformation of our function ξ with p=2 and n=1.

IV. Transformed page count based measures of word relatedness

In this section we will apply our transformation function ξ on most popular co-occurrence measures, Jaccard, Dice, Simpson and Cosines with: q=e ,p=10 and n=1/2 and e: is the Napier's constant.

Given two terms t_1 , t_2 and a similarity function Measure(t_1,t_2)>0. The general transformation formula of Measure (t_1,t_2) function to TMeasure(t_1,t_2) function is defined as:

$$TMeasure(t_1,t_2)=\begin{cases} e^{\frac{1}{2}\log_{10}(Measure(t_1,t_2))} & Measure(t_1,t_2)>0 \\ 0 & Measure(t_1,t_2)=0 \end{cases} \quad (9)$$

A. Transformed Jaccard Coefficient

We define the transformed Jaccard coefficient (TJaccard) on Jaccard [3, 8] as:

$$TJaccard(t_1,t_2)=\begin{cases} e^{\frac{1}{2}\log_{10}(\frac{count(t_1,t_2)}{count(t_1)+count(t_2)-count(t_1,t_2)})} \\ 0 \qquad count(t_1,t_2)=0 \end{cases} \quad (10)$$

B. Transformed Dice coefficient

We define the transformed Dice coefficient (TDice) on Dice coefficient [2, 3] as:

$$TDice(t_1,t_2)=\begin{cases} e^{\frac{1}{2}\log_{10}(\frac{2*count(t_1,t_2)}{count(t_1)+count(t_2)})} \\ 0 \qquad count(t_1,t_2)=0 \end{cases} \quad (11)$$

C. Transformed Simpson Coefficient

We define the transformed Simpson coefficient (TSimpson) on Simpson coefficient [2, 3] as:

$$TSimpson(t_1,t_2)=\begin{cases} e^{\frac{1}{2}\log_{10}(\frac{count(t_1,t_2)}{Min(count(t_1),count(t_2))})} \\ 0 \qquad count(t_1,t_2)=0 \end{cases} \quad (12)$$

D. Transformed Cosine Coefficient

We define the transformed cosine coefficient (TCosine) on cosine measure as:

$$TCosine\,(t_1,t_2)=\begin{cases} e^{\frac{1}{2}\log_{10}(\frac{count(t_1,t_2)}{\sqrt{count(t_1)*count(t_2)}})} \\ 0 \qquad count(t_1,t_2)=0 \end{cases} \quad (13)$$

V. SEMANTIC SIMILARITY BASED ON TITLE APPROACH

Our idea is to find an attribute that is short enough for the co-occurrence to be considered and good enough to describe document's content. In order to measure semantic similarity between two given terms t_1, t_2, the proposed approach will search for the terms t_1 and t_2 in the title of the document instead of the content of the document.

1. Search in document titles for term t_1.

   Let count (t_1), be the number of documents containing term t_1 in the title.

2. Search in document titles for term t_2.

   Let count (t_2), be the number of documents containing term t2 in the title.

3. Search in document titles for both terms t_1 and t_2.

   Let count (t_1, t_2), be the number of documents containing both terms t_1 and t2 in the title.

4. Compute scores using count (t_1), count (t_2) and count(t_1,t_2). The resulting score is a measure of similarity

In order to compute count (t_1), count (t_2) and count (t_1, t_2) our approach uses the web as a corpus and a web search engine such Google as an interface.

Web search engines index billion of pages on the web and provide an estimate of page counts as a result for a searching term. Most search engines provide an interface to search for a term or more using Boolean operators on document content. By using search engine operators, it is also possible to search for documents based on document attributes such as title, URL, Anchors, etc. For example Google provides the operator "intitle:" to search for a term in a document title and "inurl:" operator to search for a term in a document URL.

VI. EXPERIMENTS

This section introduces the evaluation of the most popular semantic similarity measure to three attributes. The attributes are the URL of the document, title of the document and the content of the documents denoted respectively as "URL", "Title" and "Doc" in our experiments. Two sets of prevalent human benchmark data are employed: Rubenstein and Goodenough (R&G) dataset [5] data set and Miller and Charles (M&C) dataset [6]. The Pearson Product-moment Correlation Coefficient [10] is employed to calculate the consistency between similarity ratings. The Pearson product-moment correlation coefficient, r_{xy}, between human ratings X and the association scores Y computed by a score measure.

$$r_{xy}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n\sigma_x\sigma_y} \quad (14)$$

Where $\bar{x}$ and $\bar{y}$ are the means of X and Y; σ_x ,σ_y are the standard deviations of X and Y respectively ; n is the total number of observations.

In our experiments we are setting N, the number of indexed pages by the search engine, to 10^{11}. The number of indexed pages is a requirement for NGD distance whereas N has no affect on PMI correlation as Pearson's correlation coefficient is invariant against linear transformations.

A. R&G dataset

R&G [5] conducted quantitative experiments with a group of 51 human judges who were asked to rate 65 pairs of English words on the scale of 0.0 to 4.0, according to their similarity of meaning [2]. A word relatedness measure is evaluated using the correlation between the relatedness scores it produces for the word pairs in the benchmark dataset and the human ratings.

“Fig. 1” reports the absolute correlation coefficients between relatedness scores on R&G dataset with human ratings for different measures and different attributes.

B. *M&C dataset*

M&C [6] repeated the same experiment of R&G [5] restricting themselves to 30 pairs from the original 65, and then obtained similarity judgments from 38 human judges [2]. Most researchers used 28 word pairs of the M&C [6] dataset because two word pairs were omitted from the earlier version of WordNet. We score the word pairs in M&C dataset using the page-count-based similarity scores defined in section 2.

“Fig. 2” reports the absolute correlation coefficients between relatedness scores with human ratings on M&C dataset for different attributes.

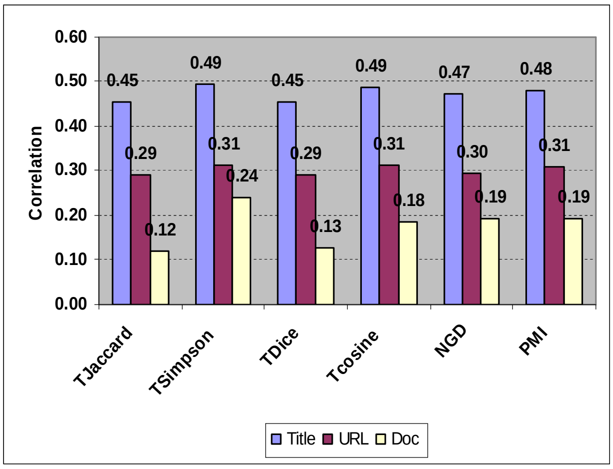

Fig. 1. Similarity correlations by attribute on R&G dataset

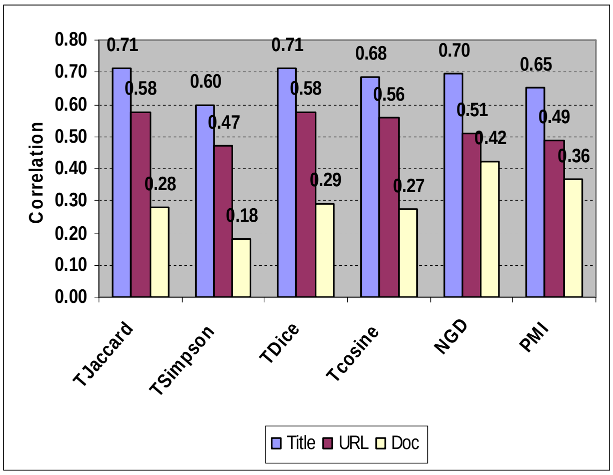

Fig. 2. Similarity correlations by attribute on M&C dataset

C. *Discussion*

All measures on R&G data-set “Fig. 1” and M&C data-set “Fig. 2” have shown a correlation increase on URL and title attributes compared to the document content.

Considering a document as a bag of word has shown bad results compared to the title and URL attributes on both datasets. The reason is the proximity or relative position of words is not considered. Based on these results, we confirm that the proximity of words improves the words relatedness. URL is performing higher than “Doc” and lowers than “Title”, for all measures on both datasets. The reason is that most users, name the URL with document’s title and this explains some of the users and software’s behaviors. For instance, Microsoft word proposes the title as a name of a file when the document is saved for the first time (unnamed document).

On M&C dataset, NGD and PMI have shown an average correlation increase of about 28.5%, whereas TJaccard, TSimpson, TDice and TCosine have shown an average correlation increase of about 42% on the title attribute, compared to document content. . On R&G dataset, NGD and PMI have shown a correlation increase of about 28.5 %, whereas TJaccard, TSimpson, TDice and TCosine have shown an average correlation increase of about 30.5% on the title attribute, compared to document content. . The correlation increase is at least of 28% on M&C dataset and 25% on R&G dataset for all measures. The title attribute has shown good results for all measures on both datasets. The reason is the context of words in the title is considered. The best results shown in our experiments are on M&C dataset for TJaccard, TSimpson and TDice measures.

Our approach outperforms similarity measures defined over snippets alone. Results reports in [4] indicate that co-occurrence Double Checking (CODC) measure [13] based on snippets alone has achieved 0.6936, whereas [12] method achieved a correlation of 0.5797 on M&C dataset. However, our approach did not outperform hybrid methods based on snippets and page counts where [4] claims an achievement of 83% on M&C dataset.

VII. RELATED WORKS

Many researchers [3, 4, 9, 11, 12 and 13] have used the web search engines to compute the semantic similarity between words. Most web search engines return page counts and snippets of a query term.

Page count of a query is an estimate of the number of pages that contain the query words. The web search engine based approaches to measure semantic similarity between words can be categorized to page count based approaches, snippets based approaches and hybrid approaches.

[11] proposed a distance metric between words using only page counts retrieved from a web search engine named Normalized Google Distance (NGD). NGD distance did not take into account the context in which the words co-occur. Our approach proposes the title of a document as an alternative in order to compute semantic similarity.

[9] defined a point-wise mutual information (PMI-IR) measure using the number of hits returned by a Web search engine to recognize synonyms. PMI-IR used AltaVista's NEAR operator to calculate page counts. "NEAR" search operator of AltaVista is an essential operator in the PMI-IR method. However, it is no longer in use in AltaVista.

Snippets were used by [12] in order to measure semantic similarity between any given two words. For each word, they collect snippets from a search engine and represent each snippet as a TF-IDF-weighted term vector. A double-checking model using text snippets returned by a Web search engine to compute semantic similarity between words has been developed by [13]. For two given words P and Q, they count the occurrences of word P in the snippets for word Q and the occurrences of word Q in the snippets for word P [3]. Snippets based approaches is an alternative to page counts approaches. However, Snippets based approaches have to deal with the extra processing coming from snippets extraction and processing. The proposed approach based on page counts alone on the title attribute has no extra processing and results are close or better than snippets based approaches.

[4] used page counts and lexical syntactic patterns extracted from snippets for measuring the semantic similarity between words. In order to compute similarity, [4] has to query the search engine to extract counts, extract snippets, process the snippets to extract pattern, build a vector space, train SVM, etc.

VIII. CONCLUSION AND FUTURE WORK

Measuring similarity between words using a search engine based on page counts alone is a challenging task. In this paper, we proposed a transformation function for web measures along with a novel approach based on page counts of document's title to measure semantic similarity between two given words. The experiments show that measuring similarity using page counts alone based on document's title achieved a correlation up to 71% on M&C data-set. Our approach outperforms similarity measures defined over snippets alone. In the future we will be employing our approach on named entities and improving similarity score results using our transformation function.

IX. REFERENCES

[1] Liu, G.; Wang, R.; Buckley, J. & Zhou, H. M. (2011), A WordNet-based Semantic Similarity Measure Enhanced by Internet-based Knowledge., in 'SEKE' , Knowledge Systems Institute Graduate School, pp. 175-178 .

[2] Aminul ISLAM Evangelos MILIOS V lado KEŠELJ. (2008). Comparing Word Relatedness Measures Based on Google n-grams. https://web.cs.dal.ca/~eem/cvWeb/pubs/2012-Aminul-Coling.pdf

[3] Bollegala, D., Matsuo, Y., and Ishizuka,M. (2007) Measuring semantic similarity between words using web search engines. In Proceedings of the 16th international conference on World Wide Web (WWW '07). ACM, New York, NY, USA, 757-766.

[4] Bollegala, D., Matsuo, Y., and Ishizuka, M. (2011). A web search engine-based approach to measure semantic similarity between words. IEEE Trans. on Knowl. and Data Eng., 23(7):977–990.

[5] Rubenstein, H. and Goodenough, J. B. (1965). Contextual correlates of synonymy. Communications of the ACM, 8(10):627–633.

[6] Miller, G. A. and Charles, W. G. (1991). Contextual correlates of semantic similarity. Language and Cognitive Processes, 6(1):1–28.

[7] Stan Matwin, Joseph De Koninck, Amir H. Razavi, and Ray Reza Amini. (2010). Classification of Dreams Using Machine Learning. In Proceedings of the 2010 conference on ECAI 2010: 19th European Conference on Artificial Intelligence, Helder Coelho, Rudi Studer, and Michael Wooldridge (Eds.). IOS Press, Amsterdam, The Netherlands, The Netherlands, 169-174.

[8] Salton, G. and McGill, M. (1983). Introduction to Modern Information Retrieval. McGraw- Hill.

[9] Turney, P. (2001). Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. In Proceedings of the Twelfth European Conference on Machine Learning (ECML 2001), pages 491–502, Freiburg, Germany.

[10] J. L. Rodgers and W. A. Nicewander (1988) "Thirteen ways to look at the correlation coefficient," The American Statistician, 42(1), pp.59–66, 1988

[11] Cilibrasi, R. L. and Vitanyi, P. M. B. (2007). The Google similarity distance. IEEE Trans. on Knowl. and Data Eng., 19(3):370–383.

[12] M. Sahami and T. Heilman. (2006). A web-based kernel function for measuring the similarity of short text snippets. In Proc. of 15th International World Wide Web Conference. 2006.

[13] H. Chen, M. Lin, and Y. Wei. (2006). Novel association measures using web search with double checking. In Proc. of the COLING/ACL 2006.

[14] Gracia, J., Trillo, R., Espinoza, M., and Mena, E. (2006). Querying the web: a multiontology disambiguation method. In Proceedings of the 6th International Conference on Web Engineering, ICWE '06, pages 241–248, New York, NY, USA. ACM.