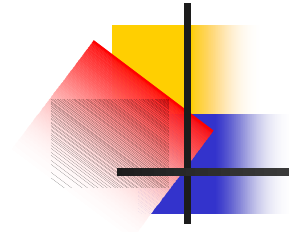


10th International Workshop on Text-based Information Retrieval
DEXA TIR 2013, Prague, Czech Republic



Word Semantic Similarity based on document title

Mohamed said Hamani saidhamani@hotmail.com

Ramdane Maamri rmaamri@yahoo.com

Presented by: Mohamed said Hamani



OUTLINE

- Approach Overview.
- Web search engine based approaches for measuring semantic similarity.
- Semantic similarity based on title approach.
- Experiments.
- Conclusion.



Approach overview

- The purpose of the paper is to measure semantic similarity between two given words based on page counts alone using a search engine as an interface and the Web as a live corpus.
- The approach exploits the titles of documents instead of the contents of documents.



Word Semantic Similarity

- Measuring the semantic similarity or dissimilarity (distance) between words is a process of quantifying the relatedness between the words using information sources [1].
- 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.



The web as Live Corpus

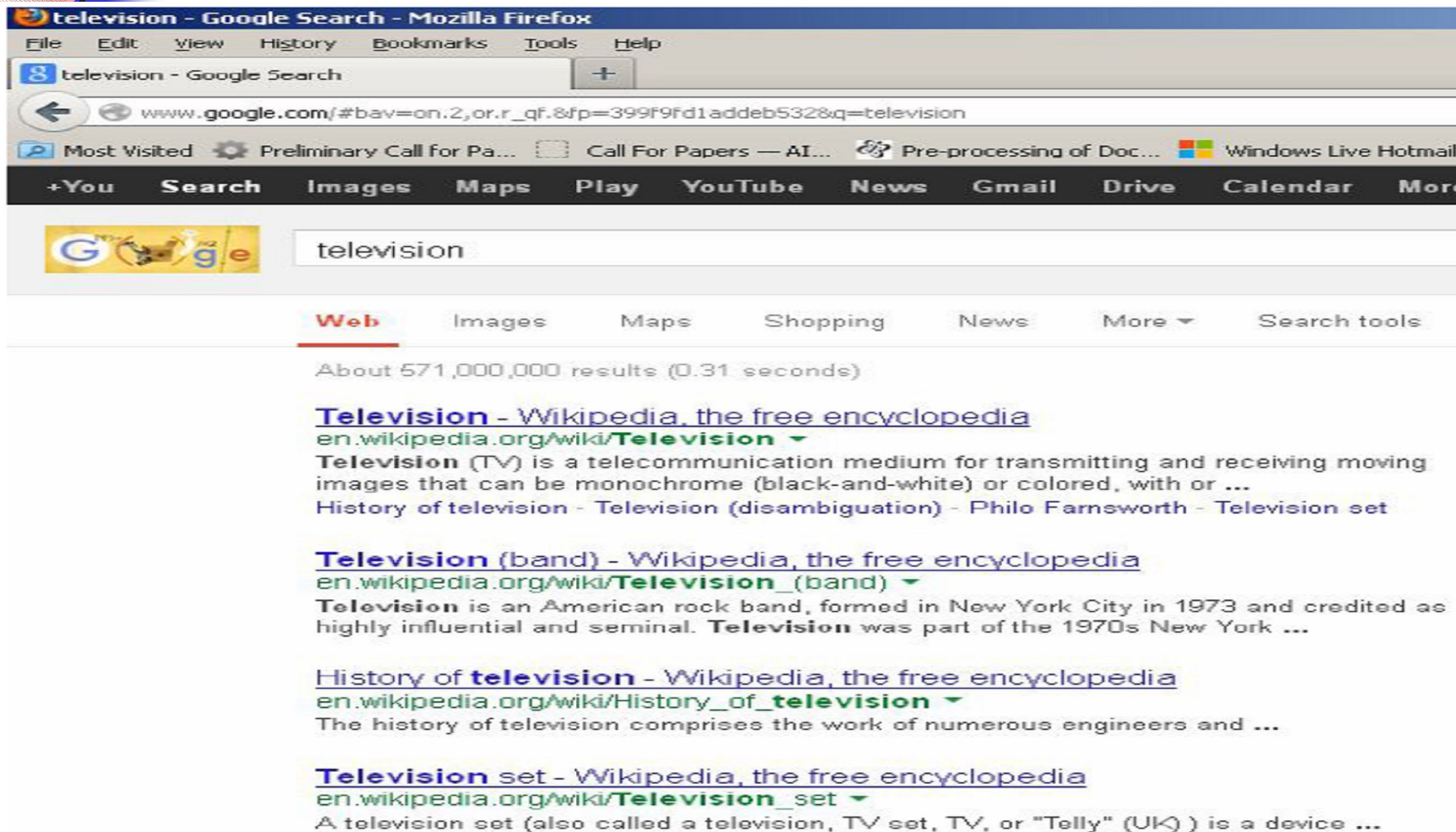
- Semantic similarity between words changes over time and across domains. New words are constantly being created as well as new senses are assigned to existing words [3].
- Manually maintaining thesauri to capture these new words and senses is costly if not impossible [3].
- Web as a live corpus instead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



Web search engines

- 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].

Example



The screenshot shows a Mozilla Firefox browser window with the title "television - Google Search - Mozilla Firefox". The address bar contains the URL "www.google.com/#bav=on:2,or:r_qf.&fp=399F9fd1addeb532&q=television". The search bar contains the text "television". Below the search bar, the "Web" tab is selected, showing search results for "television". The results include:

- [Television - Wikipedia, the free encyclopedia](http://en.wikipedia.org/wiki/Television)
en.wikipedia.org/wiki/Television
Television (TV) is a telecommunication medium for transmitting and receiving moving images that can be monochrome (black-and-white) or colored, with or ...
History of television - Television (disambiguation) - Philo Farnsworth - Television set
- [Television \(band\) - Wikipedia, the free encyclopedia](http://en.wikipedia.org/wiki/Television_(band))
en.wikipedia.org/wiki/Television_(band)
Television is an American rock band, formed in New York City in 1973 and credited as highly influential and seminal. **Television** was part of the 1970s New York ...
- [History of television - Wikipedia, the free encyclopedia](http://en.wikipedia.org/wiki/History_of_television)
en.wikipedia.org/wiki/History_of_television
The history of television comprises the work of numerous engineers and ...
- [Television set - Wikipedia, the free encyclopedia](http://en.wikipedia.org/wiki/Television_set)
en.wikipedia.org/wiki/Television_set
A television set (also called a television, TV set, TV, or "Telly" (UK)) is a device ...

Web search engine based approaches for measuring semantic similarity



- The web search engine based approaches for measuring semantic similarity between words can be categorized to:
 - Page counts based approaches [9, 11].
 - Snippets based approaches [12, 13].
 - Hybrid approaches [3, 4].



Page counts based approaches for measuring semantic similarity

- Page counts based approaches use the page counts alone returned from search engine as co-occurrence statistics to compute the semantic similarity between words.
- Drawback
 - page counts alone methods ignore word positions in a page considering the document as a bag of words, whereas “co-occurrence should be considered in a specific context or in a window of limited sizes such 3 to 7 words before or after a target word” [7].

Page count based measures of word relatedness

- There is a relatively large number of co-occurrence measures in the literature such as

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

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

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

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

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

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



The idea of our approach

- 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.
- 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 using search engine operators.
- 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.
- The paper focus on document's title and study the URL and document content as well.



Semantic similarity based on title approach

- Given two terms t_1, t_2
 - 1. Search in document titles for term t_1 .
 - Let $\text{count}(t_1)$, be the number of documents containing term t_1 in the title.
 - 2. Search in document titles for term t_2 .
 - Let $\text{count}(t_2)$, be the number of documents containing term t_2 in the title.
 - 3. Search in document titles for both terms t_1 and t_2 .
 - Let $\text{count}(t_1, t_2)$, be the number of documents containing both terms t_1 and t_2 in the title.
 - 4. Compute scores using $\text{count}(t_1)$, $\text{count}(t_2)$ and $\text{count}(t_1, t_2)$. The resulting score is a measure of similarity.



Transformed page count based measures of word relatedness

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

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

For instance the transformation of Jaccard to TJaccard is:

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

$$\text{TJaccard}(t_1, t_2) = \begin{cases} e^{\frac{1}{2} \log_{10}\left(\frac{\text{count}(t_1, t_2)}{\text{count}(t_1) + \text{count}(t_2) - \text{count}(t_1, t_2)}\right)} & \text{count}(t_1, t_2) > 0 \\ 0 & \text{count}(t_1, t_2) = 0 \end{cases}$$



Experiments

- Evaluation of the most popular semantic similarity measures 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” .
- Two sets of prevalent human benchmark data are employed:
 - Rubenstein and Goodenough (R&G) dataset [5] data set.
 - Miller and Charles (M&C) dataset [6].
- The Pearson Product-moment Correlation Coefficient [10] is employed to calculate the consistency between similarity ratings.

Similarity correlations by attribute on R&G dataset

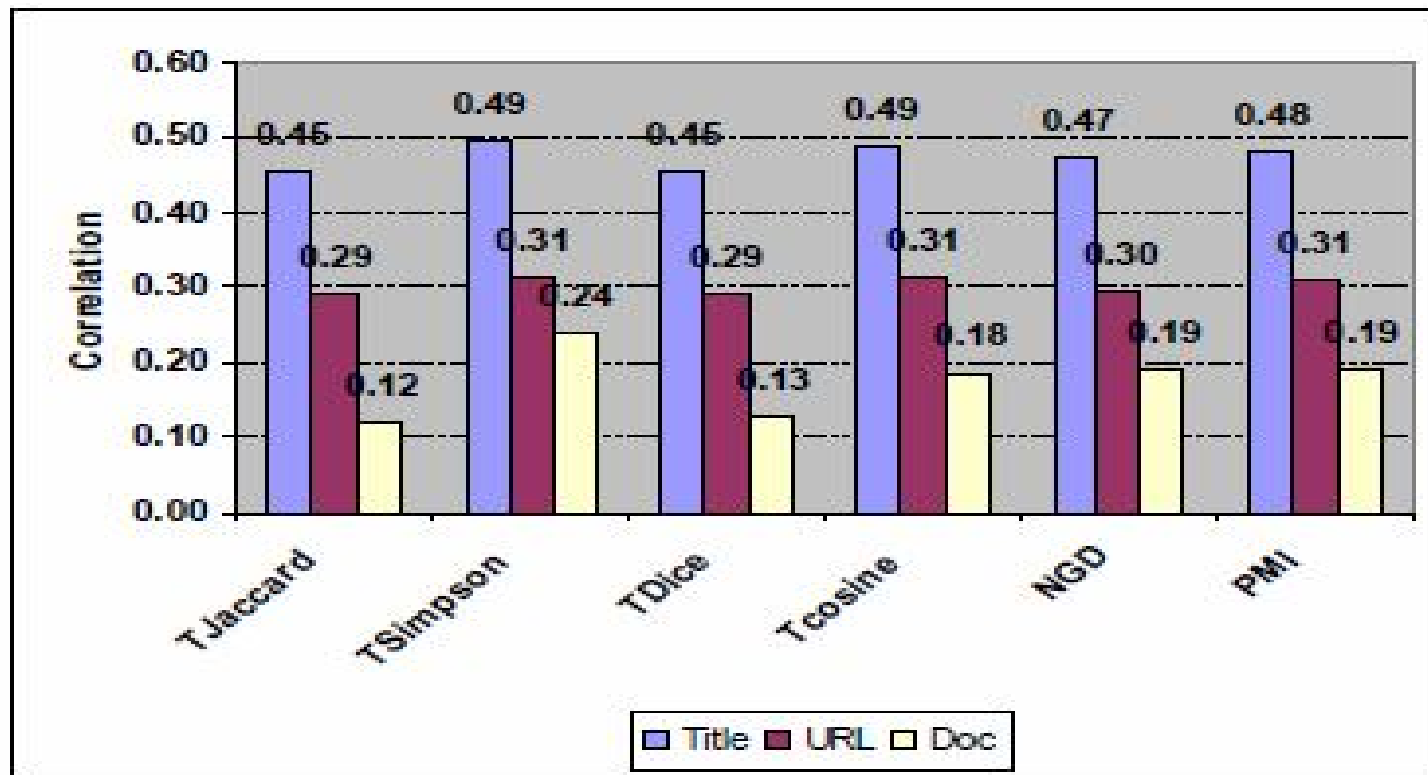


Fig. 1. Similarity correlations by attribute on R&G dataset

Similarity correlations by attribute on M&C dataset

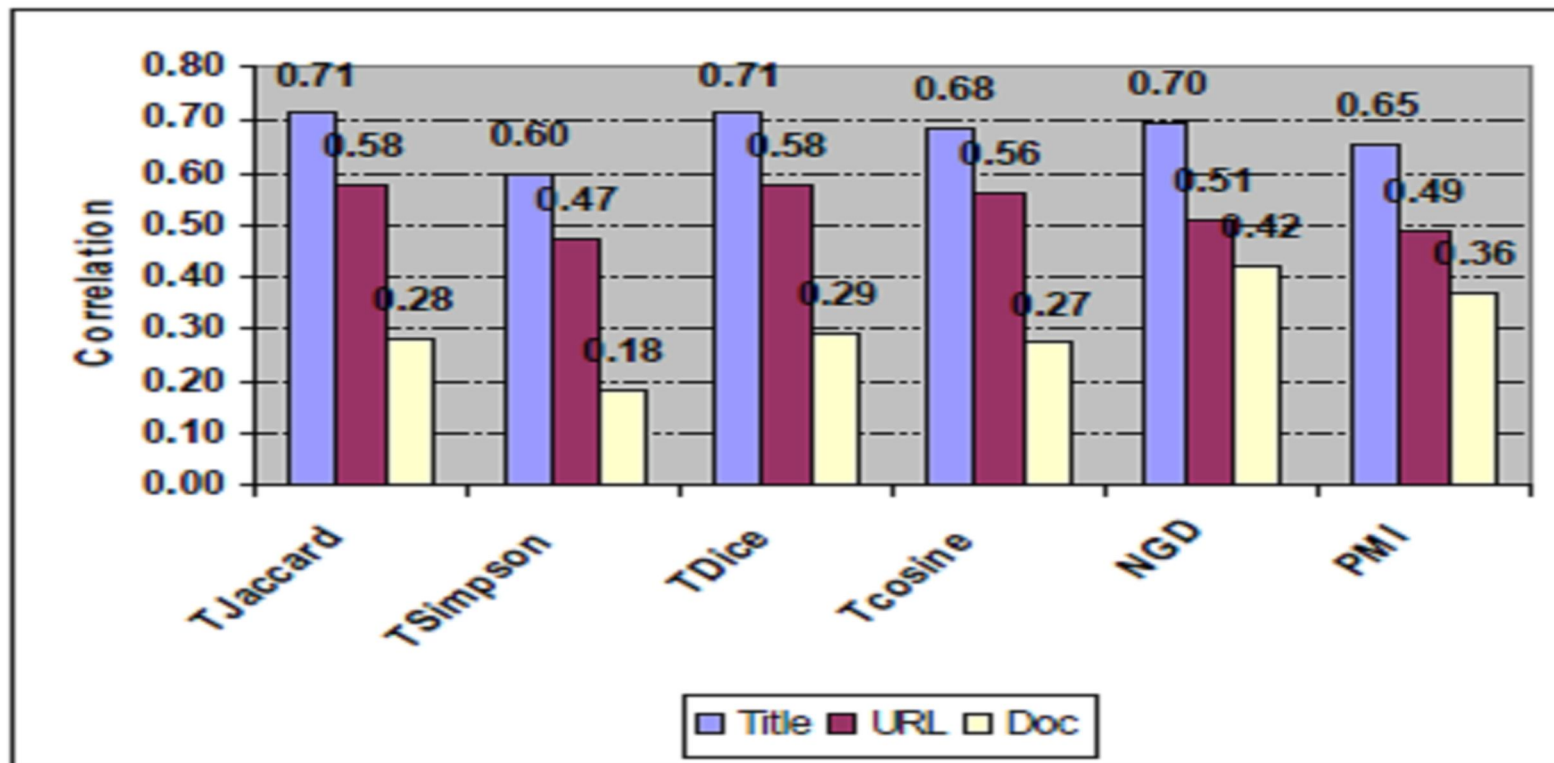
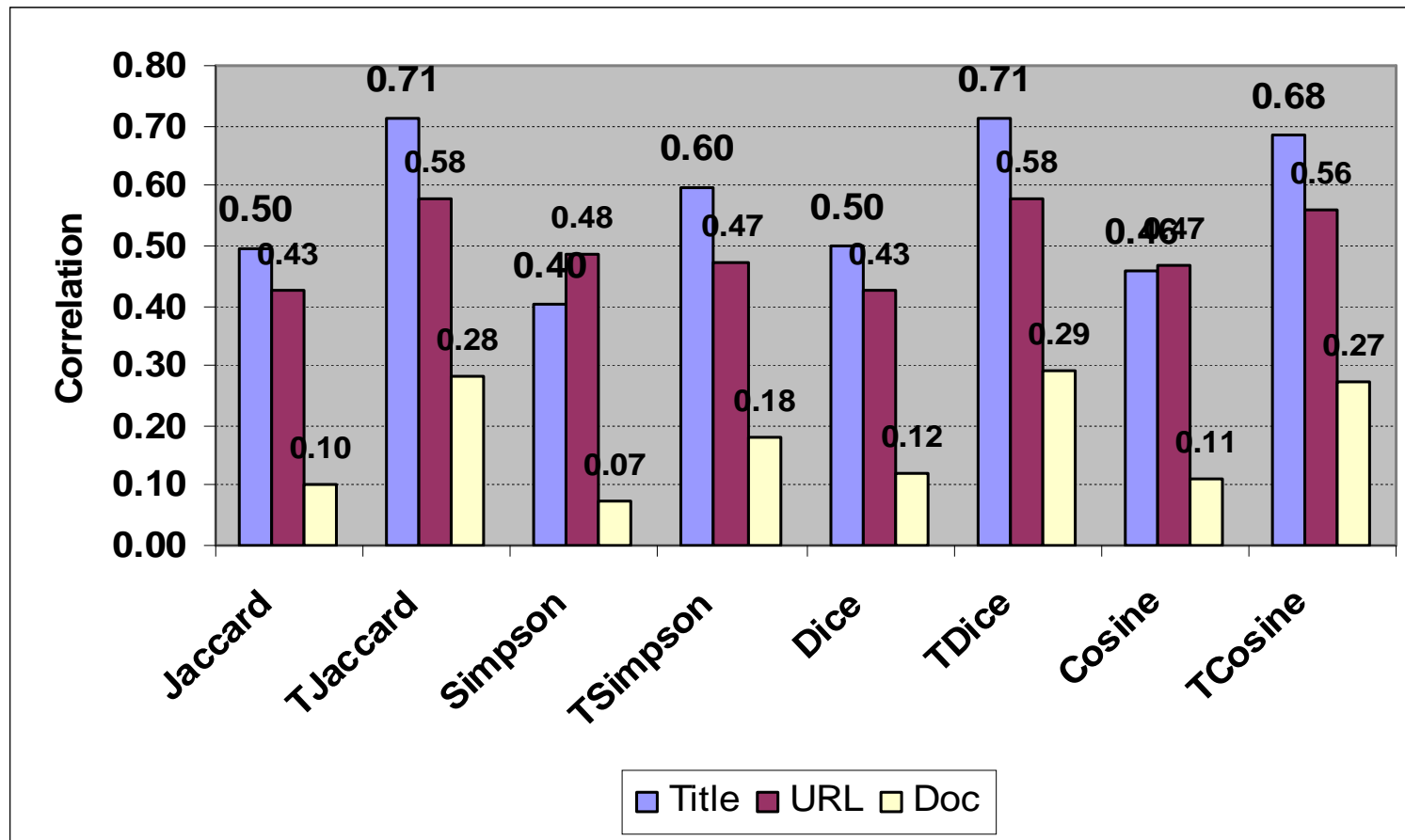


Fig. 2. Similarity correlations by attribute on M&C dataset

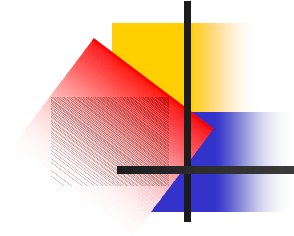
Similarity correlations by attribute on M&C dataset for Measure and TMeasure





Conclusion

- Word semantic similarity based on document title using page counts alone approach, performs better than URL and document content.
- TMeasure performs always better than Measure.
- Our approach reached 71% and outperforms similarity measures defined over snippets alone 0.58 in [12] and 0.69 in [13] based on results reported in [4].



Thank You!



References

- [1] Liu, G.; Wang, R.; Buckley, J. & Zhou, H. M. (2011), A WordNetbased Semantic Similarity Measure Enhanced by Internet-based Knowledge., in 'SEKE' , Knowledge Systems Institute Graduate School, pp. 175-178 .
- [2] Aminul ISLAM Evangelos MILIOS V Iado 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 multontology disambiguation method. In Proceedings of the 6th International Conference on Web Engineering, ICWE '06, pages 241–248, New York, NY, USA. ACM.