Comparing the accuracy of the semantic similarity provided by the Normalized Google Distance (NGD) and the Search Term Recommender (STR).

Wilko van Hoek, Brigitte Mathiak, Philipp Mayr, Sascha Schüller

GESIS – Leibniz Institute for the Social Sciences, Bonn, Germany

Extended Abstract

While Knowledge Organization Systems (KOS) support the convenience of structuring information within e.g. repositories, for users unfamiliar with a KOS, it is sometimes unclear how to obtain the appropriate search term for their needs. One method to address this problem is to determine the terms of the thesaurus that are closest to the search term in a semantic way. Therefore it is necessary to calculate the semantic similarity of the search term and the terms of the thesaurus. A commonly approach to conduct such a co-word analysis is to make use of latent semantic analysis (LSA), which relies on a statistical analysis based on word occurrences in a training corpus. We have formerly developed the system STR that is based on pLSA and support vector machines (SVM). Another approach is the NGD derived from the number of hits supported by search engines. For this paper, we will compare both methods. [01, 03, 04]

The STR, implemented in sowiport[1], is a system that interactively recommends search terms from the Thesaurus for the Social Sciences (TheSoz), which is used in sowiport. Assume a user is looking for documents related to the topic ‘evaluation’. For this terms covers different fields, the user might want to search for the term in a specific context. The STR would suggest among others the terms ‘quality assurance’, ‘methodology’ and ‘method’. As these terms are used to classify the documents in sowiport, the user is now able to refine his search by specifying the appropriate context. The basis for the recommendation in the STR is a pLSA, performed by the commercial indexing software Mindserver. The software was trained on title/abstracts from the Literature Information System for the Social Sciences (SOLIS) and the TheSoz. The training set was used to count co-occurrences of terms within the title/abstract and the terms of the controlled vocabulary. Each term gets a similarity rating between 0 (not similar) and 1 (similar). [04, 05, 06]

On contrary to the LSA our NGD does not rely on co-occurrences in the SOLIS-corpus but on webpages processed by the search engine Google. To be more specific, the numbers of websites returned, when searching for a certain term. It uses three search requests to estimate the similarity of term A and term B. The first two requests determine the number of websites returned for either the term A and the term B and the third request determines the number of websites for the term A and B. These results are then set into relation, to calculate a value of similarity of the two terms. The lower the NGD value, the closer the terms are. We changed that to the system used by the STR (1 means similar). [02, 03]

In our case study we will estimate semantic similarities using both the NGD and the STR. We aim to find out whether these similarities are in any way compatible, so that they could be used interchangeably. In a first step, we randomly extracted a subset of user-generated

[1] http://gesis.org/sowiport

search terms. These terms were logged anonymously from sowiport. After filtering less appropriate terms (terms such as 'XXETYgfLtQVfOq'), in total we obtained a set of 88 terms. We performed a comparison estimating the most accurate TheSoz terms for our subset using the STR. The maximum number of result returned from the STR in this scenario was limited to 50. The STR returned less than 50 recommendations for some of the search terms. All together we obtained 3814 pairs of terms and corresponding recommendations. Afterwards we calculated the NGD pair wise for the resulting set and evaluated statistical Measures as well as analyzed the top ten of both the STR and NGD.

The Outcome of the comparison was very heterogeneous. While in some cases the top ten recommendations of STR and the ones based NGD showed good agreement, in some other cases they were nearly disjoint. Beside the degree of accordance the accuracy of recommended terms differs too. At a first result, both seem to be more accurate in some cases and also there are cases were both recommendations are accurate, but they lack of accordance. In Table 1 and 2 we show two example results of our comparison.[2]

Search term: *immigrationpolitics*
Number of equal terms in top 10: *6*

STR-recommendation		NGD-results (sorted by value)		STR-NGD
right of asylum	0.9179	immigration	0.9717	0.0538
situation	0.8993	migration policy	0.9380	0.0387
capital flow	0.8758	Maghreb country	0.8882	0.0124
factor mobility	0.8724	Rawls, J.	0.8837	0.0112
Maghreb country	0.8266	foreign worker	0.8657	0.0390
immigration policy	0.8241	immigration policy	0.8648	0.0407
migration potential	0.8052	capital flow	0.8587	0.0534
collective consciousness	0.7964	right of asylum	0.8555	0.0591
refugee	0.7629	migration potential	0.8536	0.0907
migration policy	0.7434	employment research	0.8382	0.0948

Table 1: *this example shows a relative high degree of accordance. Both System produce similar results with a similar weighting.*

Statistically the results also showed some interesting responses. For a better comparison we chose the NGD normalizing factor *N* in a way that it minimizes the average difference between the values of NGD and STR. Summing the pairwise calculated absolute difference leads to a value of 0.36. The same sum for squared differences even was 0.48. Regarding the fact, that all values are between 0 and 1, these results confirm the above observance of heterogeneity.

As a next step we plan to calculate the NGD between a subset of user terms and all of the TheSoz-terms (7,904 terms), to provide us with a more precise basis for our comparisons. This will take some time due to the limitation of requests by Google. In addition to first case study, we will not only compare the results of the two approaches, but let the recommendations be reviewed by experts.

[2] The original studies took place in a German environment with German search terms. For this proposal they have been translated into English.

Search term: *mobbing*
Number of equal terms in top 10: *0*

STR-recommendation		NGD-results (sorted by value)		STR-NGD
working atmosphere	1.0000	education system	0.8354	0.1646
conflict behaviour	1.0000	Victim	0.8068	0.1932
workload	1.0000	Job	0.7949	0.2051
social behaviour	0.9989	labour law	0.7873	0.2116
stress	0.9986	management style	0.7836	0.2150
employment agreement	0.9919	prevention	0.7772	0.2147
firm	0.9909	hospital	0.7754	0.2155
intervention	0.9854	medical rehabilitation	0.7676	0.2178
communicative competence	0.9815	sexual harassment	0.7654	0.2161
occupational situation	0.9743	Service	0.7620	0.2123

Table 2: *this example shows two disjoint resulting sets. While both recommendations are still appropriate, the weighting is very different.*

Based on the evaluation process described in this abstract, we hope to bring more clarity into the question of comparability of different concepts calculating the semantic similarity of words. As part of our future work we will try to perform further examinations of semantic analyses based on the web. We will examine whether web based and offline computed recommendations can be combined successfully.

References

[01] Petras, V. (2006): Translating Dialects in Search: Mapping between Specialized Languages of Discourse and Documentary Languages. PhD thesis, University of California, Berkeley. http://people.ischool.berkeley.edu/~vivienp/diss/vpetras-dissertation2006-shortformat.pdf

[02] Cilibrasi, R.; Vitányi, P.M.B. (2007): The Google similarity distance. IEEE Trans. Knowledge and a Data Engineering, 19(3):370-383, 2007. http://arxiv.org/abs/cs/0412098

[03] Cilibrasi, R.; Vitányi, P.M.B. (2009): Normalized Web Distance and Word Similarity. In Proceedings of CoRR. http://arxiv.org/abs/0905.4039

[04] Mutschke, P.; Mayr, P.; Schaer, P.; Sure, Y. (2011 to appear): Science Models as Value-Added Services for Scholarly Information Systems. In: Scientometrics. http://www.ib.hu-berlin.de/~mayr/arbeiten/Science-Models-as-Value-Added-Services.pdf

[05] Hienert, D.; Schaer, P.; Schaible, J.; Mayr, P. (2011 to appear): A Novel Combined Term Suggestion Service for Domain-Specific Digital Libraries. TPDL 2011. http://www.ib.hu-berlin.de/~mayr/arbeiten/TPDL2011_final.pdf

[06] Schaer, P; Sure, Y. (2009): User interface design for search-term recommendation and interactive query expansion services. Presentation at the 8th European Networked Knowledge Organization Systems (NKOS) Workshop. http://www.comp.glam.ac.uk/pages/research/hypermedia/nkos/nkos2009/programme.html