The Normalized Freebase Distance

Fréderic Godin, Tom De Nies, Christian Beecks, Laurens De Vocht, Wesley De Neve, Erik Mannens, Thomas Seidl, and Rik Van de Walle

In: The Semantic Web: ESWC 2014 Satellite Events, 2014.

The Normalized Freebase Distance

Fréderic Godin[1], Tom De Nies[1], Christian Beecks[2], Laurens De Vocht[1], Wesley De Neve[1,3], Erik Mannens[1], Thomas Seidl[2], and Rik Van de Walle[1]

[1] {firstname.lastname}@ugent.be
Ghent University – iMinds – Multimedia Lab, Belgium
[2] {lastname}@informatik.rwth-aachen.de
RWTH Aachen University – Data Management and Data Exploration Group, Germany
[3] KAIST – IVY Lab, Republic of Korea

Abstract. In this paper, we propose the Normalized Freebase Distance (NFD), a new measure for determing semantic concept relatedness that is based on similar principles as the Normalized Web Distance (NWD). We illustrate that the NFD is more effective when comparing ambiguous concepts.

1 Introduction

In the last decade, the *Normalized Web Distance* (NWD) [1] has proven to be a simple, yet powerful measure of the semantic relatedness between two concepts. The NWD measures the semantic relatedness between two concepts in terms of their frequency of single and mutual occurrence in web pages. Essentially, two concepts appearing together on a web page reduces their NWD.

One of the most prominent instantiations of the NWD is the *Normalized Google Distance* (NGD) [2], which is based on the Google search engine. The NGD epitomizes the utilization of the NWD to web search engines, such as Google, Bing, and Yahoo. While these web search engines have the advantage that nearly all concepts can be found, they suffer from the issue of concept ambiguity: a concept that is issued as a query to the search engine can be interpreted in different ways. For example, the concept "Washington" can either refer to the state, the capital or the former president. This disambiguation is partially or completely lost when using traditional web search engines.

In this paper, we investigate how we can keep the powerful idea behind the NWD, while disambiguating the semantic meaning of concepts. Instead of textual indexes and search engines, we rely on semantic graph-structured data stores, such as DBpedia or Freebase, which can be queried unambiguously.

2 Related Work

Measuring the semantic relatedness between single language units or phrases in terms of similarity or distance has been an active research area for a long

time. Many similarity measures [3, 5, 7] are based on static semantic networks such as *WordNet*. More recent similarity metrics such as the NGD [2] and *Flickr Distance* [9] are based on dynamic repositories of user-generated content.

The idea of using graph-structured knowledge bases such as DBpedia or Freebase to calculate the distance between two disambiguated concepts has already been proposed in literature [4, 6]. However, these distance measures rely on the direct or indirect connections between two resources, and can only be calculated by using a potentially computationally expensive algorithm, such as finding the shortest path in a graph [6], or recursively calculating similarity [4].

3 The Normalized Freebase Distance

Our approach is based on the Normalized Web Distance, which is defined as:

$$NWD(x,y) = \frac{max\{logf(x), logf(y)\} - logf(x,y)}{logN - min\{logf(x), logf(y)\}} ,$$

where $f(x)$ and $f(y)$ are the numbers of web pages containing either concept x or y, $f(x,y)$ is the number of web pages containing both concepts x and y, and N is the total number of web pages. The function f thus depends on a specific search engine. As mentioned above, the best-known implementation of the NWD is the NGD. However, since recent updates by Google resulted in the removal of the + and AND operators, the implementation and thus computation of the NGD becomes infeasible. We therefore make use of the *Normalized Bing Distance* (NBD) as a representative baseline during our evaluation, since Microsoft Bing still offers this query capability.

Our approach uses the graph-structured knowledge base Freebase to calculate these values instead of a conventional web search engine. Freebase currently contains about two billion links between concepts. Consequently, a more complex approach such as searching the shortest path between two concepts would be a computationally expensive task. Therefore, we propose to make use of a similar principle as the NWD, only making use of the incoming links to a certain concept.

For two concepts x and y, we can compare the number of concepts in the dataset with links pointing to x or y separately, to the number of concepts with links pointing to both. This is similar to the page counts provided by web search engines, where a link can be seen as an occurrence on a web page. By substituting the functions $f(x)$, $f(y)$, and $f(x,y)$ as follows:

$$f(x) = number\ of\ concepts\ linking\ to\ x,$$

$$f(y) = number\ of\ concepts\ linking\ to\ y,$$

$$f(x,y) = number\ of\ concepts\ linking\ to\ x\ AND\ y,$$

we define the Normalized Freebase Distance (NFD) in a similar manner as the NWD. For our approach, we set N to the total number of concepts in Freebase.

4 Preliminary Experiments

To verify the effectiveness of the NFD, we analyze the distance matrix of a number of ambiguous examples and compare the output with the NBD. To calculate the NFD, we set up a Virtuoso SPARQL endpoint and used the Freebase RDF dump of March 16, 2014 containing over 1.9 billion triples.

The function $f(x)$ is defined as (the function $f(y)$ is defined similarly):

```
SELECT COUNT(DISTINCT ?s) WHERE { ?s ?p <x> }
```

The function $f(x, y)$ is defined as:

```
SELECT COUNT(DISTINCT ?s) WHERE { ?s ?p1 <x> . ?s ?p2 <y> }
```

The returned triples were filtered on duplicates by removing triples that used the predicates `rdf:type` and `rdfs:label`, and forced the subject to be an URI.

To evaluate the NFD, we were particularly interested in concepts that would confuse traditional search engines. To that end, we have calculated the distance between three types of fish, and the word *bass guitar*. We were particulary interested in how search engines will deal with the fish species *bass* that is contained in the word *bass guitar*.

Table 1: Distance matrices of four concepts, using the NBD (left) and the NFD (right). For Freebase entities, their unique identifier is used as label.

	Salmon	**Trout**	**Bass**	**Bass guitar**
Salmon	0	0.072	0.133	0.283
Trout	0.072	0	0.123	0.247
Bass	0.133	0.123	0	0.086
Bass g.	0.283	0.247	0.086	0

	09777	**0cqpb**	**0cqvj**	**018vs**
09777	0	0.070	0.087	0.274
0cqpb	0.070	0	0.070	0.269
0cqvj	0.087	0.070	0	0.276
018vs	0.274	0.269	0.276	0

As can be seen in Table 1, distances between the different concepts are of the same magnitude. The NBD between *salmon* and *trout* is 0.072, while the NFB between these two concepts is 0.070. Similarly, the NBD between *salmon* and *bass guitar* and the NBD between *trout* and *bass guitar* is of the same magnitude as the NFD between the aforementioned concept pairs. However, the NBD between *bass* and the three other concepts is much more different compared to the NFD. In fact, the NBD between *bass* and *trout* and the NBD between *bass* and *salmon* is much higher than the NFB between *bass* and *trout* and the NFB between *bass* and *salmon*. Likewise, the distances between *bass guitar* and the other three concepts (last row) are of the same magnitude, except for the NBD with *bass*. There, we can observe a big drop in distance, bringing *bass guitar* to the same magnitude level as the other fish. Here, the NFD captures the distances much better than NBD. We can attribute this to the fact that the NBD only relies on the occurrence of certain words and that the NBD does not take into account that concepts may consist of multiple words that can be concepts on their own.

5 Conclusions and Outlook

We have illustrated that the Normalized Freebase Distance (NFD) allows for more effective measurement of semantic concept relatedness than the Normalized Bing Distance. Additionally, the calculation of the NFD does not require to execute computationally expensive algorithms on the Freebase data set such as the shortest path algorithm.

In future research, we plan to conduct more extensive experiments, paying more detailed attention to both the effectiveness and efficiency of the proposed distance metric in a variety of use cases.

Whereas our preliminary experiments focus on the use of Freebase, one could imagine applying this principle on the scale of the entire Web. New developments in the field of Web-scale querying could make this possible in the near future [8]. That way, we could expand our NFD to a Normalized Semantic Web Distance, which would share much of the flexibility and power of the NWD, with the added benefit of semantic awareness.

Acknowledgments. The research activities in this paper were funded by Ghent University, iMinds (by the Flemish Government), the IWT Flanders, the FWO-Flanders, the European Union, and the Excellence Initiative of the German federal and state governments.

References

[1] Cilibrasi, R., Vitányi, P.M.B.: Normalized web distance and word similarity. CoRR abs/0905.4039 (2009)
[2] Cilibrasi, R.L., Vitanyi, P.M.B.: The google similarity distance. IEEE Trans. on Knowl. and Data Eng. 19(3), 370–383 (Mar 2007)
[3] Hirst, G., St-Onge, D.: Lexical chains as representations of context for the detection and correction of malapropisms. WordNet: An electronic lexical database 305, 305–332 (1998)
[4] Jeh, G., Widom, J.: Simrank: a measure of structural-context similarity. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 538–543. ACM (2002)
[5] Jiang, J.J., Conrath, D.W.: Semantic similarity based on corpus statistics and lexical taxonomy. arXiv preprint cmp-lg/9709008 (1997)
[6] Passant, A.: Measuring semantic distance on linking data and using it for resources recommendations. In: AAAI Spring Symposium: Linked Data Meets Artificial Intelligence (2010)
[7] Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. arXiv preprint cmp-lg/9511007 (1995)
[8] Verborgh, R., Vander Sande, M., Colpaert, P., Coppens, S., Mannens, E., Van de Walle, R.: Web-scale querying through linked data fragments. In: Proceedings of the 7th Workshop on Linked Data on the Web (2014)
[9] Wu, L., Hua, X.S., Yu, N., Ma, W.Y., Li, S.: Flickr distance: a relationship measure for visual concepts. Pattern Analysis and Machine Intelligence, IEEE Transactions on 34(5), 863–875 (2012)