WI-IAT 2015 The final publication is available via http://doi.ieeecomputersociety.org/10.1109/WI-IAT.2015.53

Automated calculation of term relatedness weights for semantic searches

Elizabeth-Kate Gulland
Department of Spatial Sciences
Curtin University
Perth, Western Australia
Email: e.gulland@curtin.edu.au

Simon Moncrieff
Department of Spatial Sciences
Curtin University
and CRC for Spatial Information
Perth, Western Australia
Email: s.moncrieff@curtin.edu.au

Geoff West
Department of Spatial Sciences
Curtin University
and CRC for Spatial Information
Perth, Western Australia
Email: g.west@curtin.edu.au

Abstract—**Information retrieval - finding and retrieving relevant sources of data, such as documents or geospatially located records - is a bottleneck in the process of accessing online data. Metadata describing data sources is variable in quality and quantity; textual descriptions are defined by data providers and the terminology they use will not always match search terms, particularly in fields with specialised terminology, such as health. Augmenting the original query with related terms increases the likelihood of matching to relevant metadata. Related terms can be extracted from thesaurus and term definition resources or from the Semantic Web, which defines resources and relationships between them. However, relationships between terms are complicated by multiple interpretations, often dependent upon context (for example, 'sign' may mean 'road sign' or 'medical sign', such as fever). Including the strength and/or context of a relationship in a semantic link could help narrow down extra terms to those most relevant to the query. In this paper, methods for automatically calculating the relative strength of relationships between terms were investigated and compared for general and domain-specific terms. Calculations were based on a variety of textual resources including public, crowd-sourced online sources Wikipedia and Google search engine. Measures for term relatedness in a specialist domain were tested using health as a case study. Results show promise for automatic calculation of weights between terms, which can be used to develop weighted graphs for use in semantic searches.**

I. Introduction

Matching a searcher's query with data appropriate to their needs is a significant problem for information retrieval (IR). One way to improve this is via semantic searches which match meaning rather than text patterns, for instance by comparing the original query to related terms in metadata. Determining the strength of links between terms could improve this process of matching relevant terms.

This paper outlines an initial approach to extend automatic detection of semantic links between terms by calculating the strength of the links, using a combination of public resources including Wikipedia, Google and the manually compiled WordNet. Multiple sources of information were used to test if the narrow scope of manually-produced authoritative sources such as WordNet can be augmented with the wide and expanding coverage of crowd-sourced resources. Different contexts are considered by comparing both general and biomedical terms using the same relatedness measures. These link strengths can be used to match disparate but related terms and also to determine how those terms can be grouped. For example, the medical condition 'diabetes' can refer to several more precise conditions such as type 1 or type 2 diabetes mellitus, and can be considered a type of *autoimmune disorder* (type 1), or *lifestyle disease* (type 2), amongst other groupings. A term may also be partially related to a group via a weaker relatedness link: diabetes is not considered a *genetic disorder* but there is a genetic component to its risk factors.

This first stage focusses on calculating relatedness weights between pairs of terms from general terminology resources, particularly for queries in the health domain, assuming that text-based queries for health research data can make use of general terminology, health terminology, or both. Relatedness weight measures were calculated for a combination of specialist health terms extracted from World Health Organisation information web pages, and general terms. The measures were compared, separately and collectively, to human judgements of relatedness. The strengths and weaknesses of the measures were considered in terms of their suitability for use in a system to automatically calculate initial weights between new terms in order to build and extend a weighted graph of related terms.

Within this paper, a *term* may consist of one or more *words*. A *document* is a text source such as a query string, text document, web page, database record or Wikipedia article. We introduce background into semantic search and term relatedness in section II, describe our approach in section III, and discuss results and conclusions in sections IV, V and VI.

II. Background

Early research into calculating term relatedness focussed on *similarity* of terms - 'flu' is the same as 'influenza' - or hierarchical classification - 'flu' is a type of 'disease'. Research into the broader sense of *relatedness* ('flu' is related to 'vaccine') focusses either on general terminology (e.g. [1]) or on a narrow domain such as biomedical [2], [3], [4]. These two foci have been kept separate, although many techniques are applicable across domains with relatively little adaptation [5]. Pedersen *et al* [6] investigated the application of generic measures to the biomedical domain, and research has been reported on relatedness measures separately for generic and medical domain term sets [7], [8], [9].

There has been previous work on extending traditional semantic ontologies, which define Boolean relationships between resources via triples: a subject, predicate and object ('asthma', 'is-a-type-of', 'disease'), to include definitions of the weight or probability of semantic links between terms, for example by adding an extra node to record information about the relationship between terms and including extra properties such as the probability that the link exists. It was shown that this could be used to calculate a "reliability factor" of a recorded relationship. Lacasta *et al* (2010) tested this on a lexical case study based on WordNet data and using the SKOS (Simple Knowledge Organization System) ontology [10]. Sun *et al* (2015) [11] incorporated statistical probabilities into properties in ontologies defining spatial data services.

Rules can be used to infer undefined links between concepts defined in an ontology, for instance using the Semantic Web Rule Language (SWRL). Hassanpour *et al* [12] produced an ontology of related terms based on semantic relationships calculated on the fly using multiple measures. Boolean relatedness links were recorded without maintaining ranking of results or multiple relationships between a pair of terms in different contexts.

Systems for determining biomedical term relatedness for use in semantic searches include a "conceptual framework for a Health Service Semantic Search Engine (HSSSE)" incorporating an ontology of health service knowledge, tested by medical experts [13], and a script over a medical language resource (UMLS[1]) [3].

To calculate relatedness weights between terms, the research outlined in this paper used general text resources WordNet, Wikipedia and Google. *WordNet* is a manually-produced lexical database that collects terms into synsets of similar terms, each of which can be related via semantic relations such as hyponym/hypernym: influenza (hyponym) is a type of disease (hypernym) [14]. Although it is possible to build specialised wordnets for a particular domain from a known text corpus [15], only the standard WordNet was used in this instance. The *Wikipedia* online encyclopedia contains crowd-sourced articles and as such is constantly updated and increasing in scope. It has been used to compare term relatedness by treating articles as separate text documents and/or making use of interlinks between articles [16]. Relatedness measures from Wikipedia resources have compared favourably to resources such as WordNet [17], [7]. It should be noted that the Wikipedia version used in testing can affect results which can be seen as an advantage in terms of the increasing range of possible resources with the passing of time, though it can also increase processing and/or pre-processing time. *Google*, a search engine that indexes crowd-sourced web pages, has also been heavily used as a text resource in semantic relatedness research. As with Wikipedia, the pages that Google indexes are continuously increasing so this is a resource that is updated and expanded over time.

Terminology resources in the field of health include textual dictionaries and thesauri as well as structured digital sources such as databases, XML documents and ontologies. Systems in use world-wide include ICD-10[2], SNOMED-CT[3], MeSH[4], and UMLS. Although there are a number of structured resources for comparing terminology that are specific to the biomedical domain, this is not commonly the case in other domains [5]. As a consequence, biomedical semantic search tools can have a narrow focus rather than being easily extensible to other areas that can also affect health decisions.

Individual relatedness measures investigated in this research include the Normalised Google Distance (NGD) and Wikipedia Link Measure (WLM). The *NGD* measure, as defined by Cilibrasi and Vitányi (2007) [18], estimates the distance between two terms, based on the count of web pages containing the terms. This measure is symmetrical and in the range 0-1, although use of Google hitcount estimates can cause erroneous values outside this range which, in this research, are truncated to 0 or 1 as appropriate. NGD values are inverted (i.e. 1-NGD) to convert from a distance to a relatedness measure. *WLM*, as defined by Milne and Witten (2008, 2013) [17], [19], is a relatedness weight 0-1 between two terms that are each the topic of a Wikipedia article, calculated using the number of links between articles.

Evaluating relatedness measures or methods solely against a unique dataset is useful as an initial exploration but makes different methods difficult to compare. Using consistent evaluation sets of paired terms improves the robustness of comparison tests [20]. Manually collected test sets of terms with human judgements tend to be fairly small, particularly when multiple human annotators are used, due to the costs of manual production of these resources. Examples typically include sets of word pairs with similarity or relatedness values, such as wordsim252 [21] using WordNet terms, and others [1], [5], [8]. Consider also surveys of semantic relatedness measures including Budanitsky & Hirst [1], who compared semantic relatedness algorithms against the WordNet dataset; Milne & Witten [17] and Gabrilovich & Markovitch [7] who tested and compared relatedness algorithms using Wikipedia and other knowledge bases including WordNet; and Zhang *et al*'s [5] 2012 survey of semantic relatedness methods and results. Annotated word sets in specialised domains have also been developed and tested, including biomedical sets such as MESH36 [8], Ped30 [6] from SNOMED-CT terms, and Pak587 and Pak101 [22], [2] from UMLS terms. In each of these cases, the number in the name refers to the number of term pairs that were annotated by one or more expert users.

[1] Unified Medical Language System http://www.nlm.nih.gov/research/umls/

[2] International Classification of Diseases and Related Health Problems http://apps.who.int/classifications/icd10/browse/2014/en

[3] Systematized Nomenclature of Medical Terms - Clinical Terms http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html

[4] Medical Subject Headings http://www.nlm.nih.gov/mesh/meshhome.html

III. Approach

A set of 50 terms, henceforth called *mixmed50*, was produced by extracting health terms from topics in online World Health Organisation information pages[5] that matched to articles in a simple English Wikipedia dump file[6]. General terms were selected from linked articles such as 'poodle' from 'diabetes' → 'poodle' (full English Wikipedia) and 'plow' from 'vaccination' → 'cattle' → 'plow' (via the *suggest* service of a Wikipedia Miner (WM) toolkit instance on the Wikipedia dump file [19]). Because all mixmed50 terms match to a Wikipedia article, the set lacks direct synonyms such as 'flu' and 'influenza'. Each term in this set of 50 terms was manually compared to all others and given approximate relatedness scores between 0 (unrelated) and 1 (strongly related, though not necessarily synonymous). These manual relatedness measures were used as a ground truth to compare with calculated weights.

The existing Ped30 set of 50 biomedical term pairs was also tested for comparison with tests by previous researchers, after scaling the human judgements to the 0-1 range to be consistent with mixmed50 results.

A set of statistical measures were recorded for each term pair: 1-NGD, WLM, and WordNet connectivity (WN). 1-NGD was calculated for paired terms using hitcounts extracted directly from calls to the Google website, and was also calculated for hitcounts filtered by .gov (government) and .org (organisation) website domains, to test if these theoretically more authoritative websites would affect the accuracy of the distance measure. WLM was calculated between all paired terms using the *compare* service of the WM instance described above. Where one or more terms in a pair could not be matched directly to an article in the WM, a WLM value of 0 was assigned. WN was calculated using WordNet version 3.0 via the Python nltk library. Each term in a pair x,y was matched to a synset X,Y. If X and Y were the same or a synonym or hypernym/hyponym relationship existed between them, a weight of 1.0 was assigned. If the relationship was in one direction only (X→Y OR Y→X), a weight of 0.5 was assigned. Synsets were tested to a maximum of two links, e.g. X → A → Y. All other pairs of terms were assigned a weight of 0, including where no synset was found for a term. Each measure was normalised to the 0-1 range and then combined into a single measure using an L2 Frobernius norm.

IV. Results

Relatedness values (0-1) were evaluated for WLM, WN, 1-NGD and the combined measure in comparison to the ground truth, which was taken as a human judgement of at least 50% relatedness. For each calculated measure, paired terms were classified as related if the measure was over a threshold value from [0,0.1,0.2,...0.9]. A Receiver Operating Characteristic (ROC) plot of true positive versus false positive rates at each threshold value, where points closest to FPR=0 and TPR=1 are the most desirable, is shown in Figure 1.

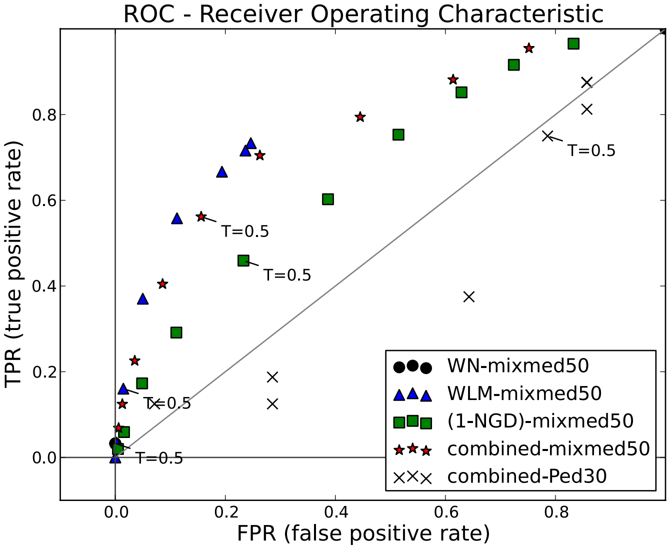

Fig. 1. Relatedness matches with agreement thresholds (T) from 0 to 0.9

Results shown are for unfiltered Google searches. Incorporating 1-NGD values for .gov and .org web domains into the combined measure resulted in values closer to the neutral diagonal line and had lower precision and accuracy, so appeared to decrease the effectiveness of the combined measure.

WordNet results were clustered around maximum or minimum TPR+FPR, depending upon the threshold value, as few of the terms were represented in WordNet.

At a threshold of 0.4, WLM had an accuracy of 0.78 and a precision of 0.71. The combined measure at the same threshold was less reliable, with an accuracy of 0.72 and precision of 0.58. The Ped30 set of pre-annotated biomedical term pairs had an accuracy of 0.5 and precision of 0.52 at this threshold.

V. Discussion

WLM gave the most reliable results of the three measures for the mixmed50 testset but performed poorly on the Ped30 set of biomedical terms. Similarly, WordNet accuracy was lower for Ped30 than mixmed50, potentially because of the specialised domain of many of the terms in both sets, particularly Ped30. It is theorised that a test set including more alternative spellings and synonyms, such as can be expected from queries from multiple users, could benefit more from the use of general thesaurus resources such as WordNet in the final distance measure. Use of a specific health text wordnet, either a pre-existing health wordnet or built upon a corpus of health-related documents, is likely to improve the rate of WordNet matches. Use of a more complete and current Wikipedia is also likely to improve WLM reliability.

As Google indexes web pages beyond the scope of Wikipedia articles, the NGD has the potential for greater coverage of terms than WLM, although it is worth noting that the NGD measure is more reliable with higher-hitcount

[5] http://www.who.int/topics/en/

[6] http://dumps.wikimedia.org/simplewiki/, 25/4/2014

terms such as returned for more common terms. Usage limits on Google searches and the lack of tools to return hitcounts, rather than search results, is a possible limitation on the use of NGD. However, it could be applied as an initial test value on whether to proceed for links between terms that cannot be matched to a Wikipedia article.

Combining relatedness measures has the potential to even out the strengths and weaknesses of individual measures, and was a reasonable starting point for a relatedness weight between terms, particularly general terms, given the lower precision and accuracy for the more specialised Ped30 set. The combined measure could be extended to include weights calculated from context-specific resources such as ICD-10 and MeSH for health. Further testing on larger biomedical term sets such as Pak101 and Mesh36 as well as sets containing more synonym-rich term pairs, such as wordsim252, is recommended to further test the use of combined measures as a temporary weight value where one or more individual measures perform poorly. As existing test sets of terms are either general or in one specialist domain (such as health), the development of larger mixed-domain sets may also be required to facilitate future experimentation into proportions of individual measures in relatedness weights under general, specialist or combined terminology scenarios.

VI. Conclusions

Flexibility of relatedness measures between terms is important for automated search solutions where queries can include both general and domain-specific terms. This research tested combinations of general term relatedness measures (based on Wikipedia, Google and WordNet) against sets of paired term in a specialist domain (health) and a mixture of general and health terms: Ped30 and mixmed50 test sets, respectively. The WLM and combined measure were more accurate for the mixmed50 set, which contained fewer highly-specialised terms than the Ped30 set. A more comprehensive Wikipedia version would improve accuracy of WLM but specialist terms are still likely to result in fewer Wikipedia article matches.

The relatedness weights could be used to build an expanded set of terms from a query to increase the chances of matching to relevant metadata. Further work on incorporating more specialised relatedness measures into the combined weight measure is needed, and more mixed-domain test sets could further facilitate this testing.

Acknowledgements

The research reported in this paper is supported by the Australian Primary Health Care Research Institute (APHCRI), which is supported by a grant from the Australian Government Department of Health. The information and opinions contained in it do not necessarily reflect the views or policy of the Australian Primary Health Care Research Institute or the Australian Government Department of Health. This work has been supported by the Cooperative Research Centre (CRC) for Spatial Information, whose activities are funded by the Australian Commonwealths Cooperative Research Centres Programme.

References

[1] A. Budanitsky and G. Hirst, "Evaluating WordNet-based measures of lexical semantic relatedness," *Computational Linguistics*, vol. 32, no. 1, pp. 13–47, 2006.

[2] S. Pakhomov, T. Pedersen, B. McInnes, G. B. Melton, A. Ruggieri, and C. G. Chute, "Towards a framework for developing semantic relatedness reference standards," *Journal of Biomedical Informatics*, vol. 44, no. 2, pp. 251–265, 2011.

[3] B. McInnes, T. Pedersen, S. Pakhomov, Y. Liu, and G. Melton-Meaux, "UMLS::Similarity: Measuring the relatedness and similarity of biomedical concepts," in *Proceedings of the 2013 NAACL HLT Demonstration Session*, 2013, pp. 28–31.

[4] V. N. Garla and C. Brandt, "Semantic similarity in the biomedical domain: an evaluation across knowledge sources," *BMC Bioinformatics*, vol. 13, no. 261, pp. 1–13, 2012.

[5] Z. Zhang, A. L. Gentile, and F. Ciravegna, "Recent advances in methods of lexical semantic relatedness - a survey," *Natural Language Engineering*, vol. 19, no. 4, pp. 411–479, 2012.

[6] T. Pedersen, S. V. S. Pakhomov, S. Patwardhan, and C. G. Chute, "Measures of semantic similarity and relatedness in the biomedical domain," *Journal of Biomedical Informatics*, vol. 40, pp. 288–299, 2007.

[7] E. Gabrilovich and S. Markovitch, "Wikipedia-based semantic interpretation for natural language processing," *Journal of Artificial Intelligence Research*, vol. 34, pp. 443–498, 2009.

[8] E. G. M. Petrakis, G. Varelas, A. Hliaoutakis, and P. Raftopoulou, "Design and evaluation of semantic similarity measures for concepts stemming from the same or different ontologies," in *4th Workshop on Multimedia Semantics*, 2006, pp. 44–52.

[9] A. Q. Li, A. Ahmed, S. Ravi, and A. J. Smola, "Reducing the sampling complexity of topic models," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, New York USA, August 24-27*, 2014, pp. 891–900.

[10] J. Lacasta, J. Nogueras-Iso, and F. J. Zarazaga-Soria, *Terminological Ontologies: Design, Management and Practical Applications*, ser. Semantic Web and Beyond. Springer eBooks, 2010, vol. 9.

[11] S. Sun, L. Wang, R. Ranjan, and A. Wu, "Semantic analysis and retrieval of spatial data based on the uncertain ontology model in digital earth," *International Journal of Digital Earth*, vol. 8, no. 1, pp. 1–14, 2015.

[12] S. Hassanpour, M. J. O'Connor, and A. K. Das, "Clustering rule bases using ontology-based similarity measures," *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 25, 2014.

[13] H. Dong, F. K. Hussain, and E. Chang, "A framework enabling semantic search in health service ecosystems," in *6th International Conference on Semantics, Knowledge and Grids*. IEEE Computer Society, 1-3 November 2010, pp. 235–242.

[14] C. Fellbaum, Ed., *WordNet: an electronic lexical database*. MIT Press, 1998.

[15] Z. Kozareva and E. Hovy, "Tailoring the automated construction of large-scale taxonomies using the web," *Language Resources and Evaluation*, vol. 47, no. 3, pp. 859–890, 2013.

[16] O. Medelyan, D. Milne, C. Legg, and I. H. Witten, "Mining meaning from Wikipedia," *International Journal of Human-Computer Studies*, vol. 67, no. 9, pp. 716–754, 2009.

[17] D. Milne and I. H. Witten, "An effective, low-cost measure of semantic relatedness obtained from Wikipedia links," in *Wikipedia and artificial intelligence: an evolving synergy*, R. Bunescu, E. Gabrilovich, and R. Mihalcea, Eds. AAAI Press, 2008, pp. 25–30.

[18] R. L. Cilibrasi and P. M. B. Vitányi, "The Google Similarity Distance," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 3, pp. 370–383, 2007.

[19] D. Milne and I. H. Witten, "An open-source toolkit for mining Wikipedia," *Artificial Intelligence*, vol. 194, p. 18, 2013.

[20] H. Dong, F. K. Hussain, and E. Chang, "A survey in semantic search technologies," in *2nd IEEE International Conference on Digital Ecosystems and Technologies*. IEEE, 26-29 February 2008, pp. 403–408.

[21] E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Pasca, and A. Soroa, "A study on similarity and relatedness using distributional and WordNet-based approaches," in *Proceedings of North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT), Colorado USA, May 31-June 5*, 2009.

[22] S. V. S. Pakhomov, B. McInnes, T. Adam, Y. Liu, T. Pedersen, and G. B. Melton, "Semantic similarity and relatedness between clinical terms: An experimental study," in *AMIA Annual Symposium*, 2010, pp. 572–576.