CO-122

USER GENERATED CONTENT AND FORMAL DATA SOURCES FOR INTEGRATING GEOSPATIAL DATA

AL BAKRI M., FAIRBAIRN D.
Newcastle University, NEWCASTLE UPON TYNE, UNITED KINGDOM

ABSTRACT

Today, problems of spatial data integration have been further complicated by the rapid development in communication technologies and the increasing amount of available data sources on the World Wide Web. Thus, web-based geospatial data sources can be managed by different communities and the data themselves can vary in respect to quality, coverage, and purpose. Integrating such multiple geospatial datasets remains a challenge for geospatial data consumers. This paper concentrates on the integration of geometric and classification schemes for official data, such as Ordnance Survey (OS) national mapping data, with volunteered geographic information (VGI) data, such as the data derived from the OpenStreetMap (OSM) project. Useful descriptions of geometric accuracy assessment (positional accuracy and shape fidelity) have been obtained. Semantic similarity testing covered feature classification, in effect comparing possible categories (legend classes) and actual attributes attached to features. The model involves 'tokenization' to search for common roots of words, and the feature classifications have been modelled as an XML schema labelled rooted tree for hierarchical analysis. The semantic similarity was measured using the WordNet::Similarity package. Among several proposed semantic similarity methods in WordNet::Similarity, the Lin approach has been adopted to give normalised comparison scores. The results reveal poor correspondence in the geometric and semantics integration of OS and OSM.

1. INTRODUCTION

Spatial data integration has been defined by Uitermark et al. (1999) as the process of establishing relationships between corresponding objects in different geospatial data sets of the same space. However, this process is far from straightforward as it must consider the many varying characteristics associated with spatial data itself. In the spatial domain, the complexity and diversity in the way the data has been captured and stored are major issues for successful data integration.

As the web becomes the dominant source of geospatial data exchange, the general public is able to participate in geospatial data collection projects, such as OpenStreetMap (OSM), to obtain maps, alternative to official data, and also to contribute to their production. OSM, started by Steve Coast in 2004, aims to build a free geographic database of the world (Ramm et al., 2011). In response, some governmental mapping agencies, such as Ordnance Survey of Great Britain (OS), now supply data through web services and are developing further functionality. One of the main products of OS is OS MasterMap for the whole of Great Britain. This is a large-scale digital map including topographic information on landscape features. It contains four layers: the topographic layer, the address layer, the integrated transport layer and the imagery layer (OrdnanceSurvey, 2010).

The diversity of available data sources on the web service provides a chance of integration in order to gain the relative benefits of each data set. For the project described here, which addresses the viability of such integration, the term 'data integration' is described following Butenuth et al. (2007). They reported that data integration is not just overlaying data sets in a geographic information system, but also assessing how well the geometric and semantic properties of one data set can be transferred to the other. In two separate databases, however, the problem of heterogeneity in geometric and in semantic aspects may lead to difficulty in attempting integration.

This paper firstly illustrates geometric similarity for formal data, such as Ordnance Survey (OS), and crowdsourced data, such as OpenStreetMap (OSM) information, with the intention of assessing possible integration. Secondly, we focus on semantic similarity of feature classification of the formal data, and the crowdsourced data. The third section discusses the semantic similarity tests in detail; this includes an introduction about WordNet database and WordNet::Similarity software, pre-processing phase and semantic similarity measurement phase, before final concluding remarks in section four.

2. GEOMETRIC SIMILARITY

In order for data set convergence, and ultimately data integration, to become useful (to assist in the development of Spatial Data Infrastructures (SDI) for example), both the geometric and semantic correspondences have to be assessed. The former relies on an assessment of geometric accuracy, an area of significant long-standing interest in the fields of surveying and geomatics. In the context of Geographic

Information Science (GIS) two possible components of geometric accuracy, positional and shape accuracy, can be considered. Positional accuracy can be defined as a measure of the difference between the position of a distinct object as recorded in the database, and its true location on the ground (Goodchild and Hunter, 1997). Shape accuracy can be examined by studying relative metrics of polygonal shape or line/boundary curvature between features.

This study has considered and reported on geometric accuracy and integration of formal and informal data of variable levels of precision (Al-Bakri and Fairbairn, 2010), by assessing the comparative positional and shape quality for Ordnance Survey (OS) and OpenStreetMap (OSM) information. Standard high quality field survey (FS) was used to create a definitive reference data set. This accurate data set was produced using the highest precision survey instruments, the data from which was used as the benchmark for formal data (OS) and User Generated Content (UGC) data (OSM) comparisons. OS, OSM and FS reference data sets were compared in two contrasting case study areas: an urban area (the town of Cramlington) and an open peri-village/rural landscape (close to the area of Clara Vale), both in Northumberland, UK.

The assessment of such positional accuracy can be facilitated by adopting established standards for spatial data accuracy, helping the measurement and reporting of accuracy across a whole dataset. Several different positional spatial data accuracy standards have been developed, such as the National Standard for Spatial Data Accuracy (NSSDA) (Congalton and Green, 2009). NSSDA specifies that the positional accuracy can be reported at ground scale rather than map scale, allowing it to be used with digital map data as well as hard copy maps. Furthermore, the NSSDA provides a formal approach to how the tested points should be identified, measured and distributed across the map. It suggests that twenty or more test points covering the research area are required to effectively test data accuracy. These points must be well defined, easy to measure and found in both tested and reference data sets. The ideal distribution of tested points is even with at least 20 percent of the points in each quadrant when the data set covers a rectangular area. The intervals between points should be at least 10 percent of the diagonal distance of the total area of the data set. The NSSDA uses Root Mean Square Error (RMSE) to estimate the positional accuracy.

Linear accuracy can be examined using measures of the shape or curvature similarity between two lines, as well as positional accuracy of points along the line. When line features such as roads or railways are considered, comparison using point accuracy is insufficient to capture the geospatial complexity of linear features. While point measures may be straightforward to understand and calculate, they do not capture all aspects of line accuracy. In research on line accuracy measurements, the buffering method and an assessment of buffer overlay, has been a popular technique (Tveite, 1999; Goodchild and Hunter, 1997). This is an iterative approach because it will be impossible to estimate the appropriate buffer size in advance. The process of gradually increasing buffer size should be terminated when the results of measuring displacement or overlap seem to stabilize.

For area shape similarity, many comparative descriptors have been studied and applied in practice. Examples include compactness (Austin, 1984), elongation (Stojmenovic and unic, 2008), convexity (Esa et al., 2006) and concavity (Ebdon, 1985). Methods that use these geometric dimensions to assess shape quality are not computationally complex and do not tell much about shape, especially irregular shapes (Ali, 2002). A more precise and useful alternative method for analyzing shapes involves moment invariant analysis, such as Hu's (1962) invariant moments, and Chen's (1993) improved moment invariants. To measure shape discrepancies between features, we use techniques based on invariant moments. Moments were first used for mechanics purposes other than shape description. Hu (1962) was the first to set out the mathematical foundation for two dimensional moments invariant and demonstrated their application to shape recognition. He proved that a proper combination of moments can provide translation, scale, and rotation invariant quantities. After Hu, several studies have explored further methods to compute moments invariant. In 1993, Chen published a paper introducing a convenient procedure to calculate the moments invariant along an object boundary. These moments are called improved moments invariant and are a reformation of Hu's moments.

The results of geometric similarity showed that the accuracy of OS data is very close to the reference FS data set. However, the positional and shape accuracy of OSM data does not match the reference or the OS data sets. The results revealed the poor positional and shape accuracy of OSM data, notably that data which can be regarded as 'soft' detail (natural features, objects with 'fuzzy' boundaries, imprecise intersections). Whilst Haklay (2010) has used an alternative approach using 'completeness' as an accuracy metric, and concluded that a comparison of OSM data with small-scale OS data is favourable, Al-Bakri and Fairbairn's (2010) emphasis on the positional accuracy of large-scale data has concluded that there are significant shortcomings in the OSM data.

3. SEMANTIC SIMILARITY

In addition to geometric accuracy, semantic similarity is another fundamental notion which needs to be addressed in the context of data integration in GIScience. This part of the study focuses on the integration of classification schemes for official data, such as Ordnance Survey (OS) national mapping data, and for VGI data, such as the data derived from the OpenStreetMap (OSM) project. A model is developed to allow legend category schemes to be submitted to standard similarity assessment software (WordNet::Similarity), a choice of approach is made and some comparison scores are presented.

3.1 WordNet::Similarity Methods and Their Applications

WordNet is a lexical on-line database for the English language, created and maintained at Princeton University, and designed to establish the connections between four types of Parts of Speech (POS): noun, verb, adjective, and adverb. These sets of words are organised into sets of cognitive synonyms ('synsets'), which represent a specific meaning of the word. A synset also has a short definition or description of the real world concept known as a 'gloss'. WordNet is particularly suitable for semantic similarity measures, as it organises the information into meaning-rich hierarchies. A component package, the WordNet::Similarity software, can be used to measure semantic similarity between a pair of concepts based on the lexical database WordNet.

WordNet::Similarity provides six measures of semantic similarity, and three measures of relatedness (Pedersen et al., (2004). Three of the six measures of similarity are based on information content of the least common subsume (LCS). These measures are res (Resnik, 1995), lin (Lin, 1998), and jcn (Jiang and Conrath, 1997). The remaining three similarity measures are based on path length methods: lch (Leacock and Chodorow, 1998), wup (Wu and Palmer, 1994), and path. The relatedness methods available are hso (Hirst and Onge, 1998), lesk (Banerjee and Pedersen, 2003), and vector (Patwardhan et al., 2003). For semantic similarity in this study, Lin's (1998) approach has been adopted because it emphasises the meaning relationship, and the range of similarity scores is between 0 and 1, thus a normalizing process is not required (unlike the Resnik method).

3.2 Semantic Similarity Measurement between Feature Classifications

Before computing the semantic similarity between corresponding features, it is necessary to split a geospatial data category name that is composed of multiple words into 'tokens' (an identifiable word or root): this operation is called 'tokenization' (Tansalarak and Claypool, 2007). As a second step of pre-processing a set of class names is used to encode feature classifications as an XML schema. Such an action allows for individual comparison of classes along with a consideration of hierarchical organisations of the entire classification array.

The feature classifications to be compared - the topographic layer of OS MasterMap and the OSM data set - each covering the same areas in both Cramlington and in Clara Vale, have been modelled as labelled rooted trees in an XML schema model. Each element or attribute of the schema is translated into a node. The semantic models of two XML schemas for part of features in Clara Vale are shown in Figure 1. Many features in each data set share common characteristics, but differences occur in the definitions of concepts, categories and classifications in each of the data sets. It can be seen from Figure 1 that there are some concepts such as 'track' that have the same meaning and the same level in both schemas. For the term 'path' in the OS model, there is no exact equivalent term in the OSM schema, although there is a synonym concept, 'footway'. For some features in OSM (e.g. 'playground' or 'garden') there is no correspondence in OS data, and vice versa. Some OS categories are combined area classes, with mixed vegetation cover defined as a sequence of terms. The use of semantic similarity for assessing common overlap for different datasets could assist in the integration of classification schemes for different sources' data sets, although this does not initially look promising from the variability shown in Figure 1.

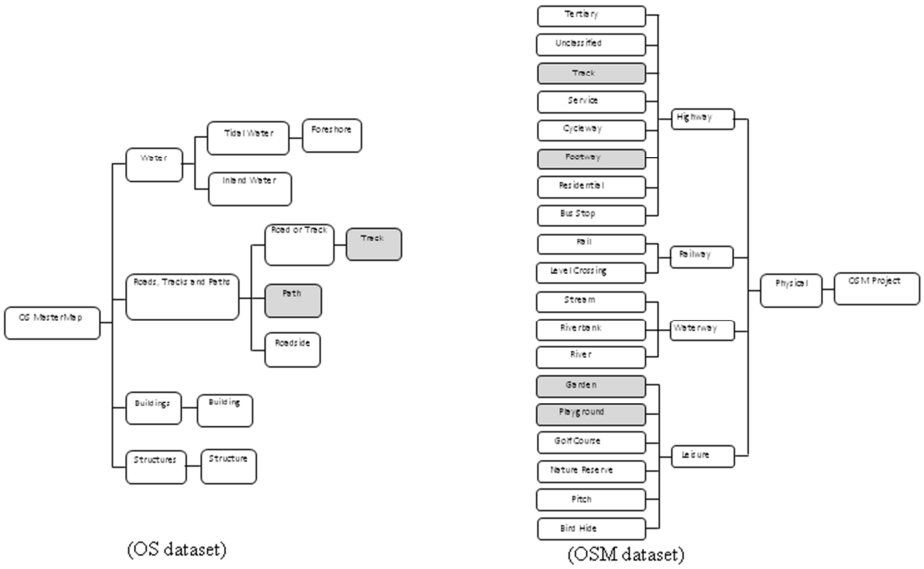

Figure 1. Part of XML schemas for feature classifications

Each named class has been tokenized into a list of words, and the WordNet::Similarity software package has been used to compute the similarity between the words. Measuring the name similarity of two sets of tokens L1 and L2 can assist in determining how linguistically close the names of two features are. The name similarity between the two sets of name tokens can be determined as the average best similarity measure for each source token with a target token, as follows (Tansalarak and Claypool, 2007):

$$N_{sim}(N_1, N_2) = \frac{\sum_{l_1 \in L_1}^{\square}[max_{l_2 \in L_2} sim(l_1, l_2)] + \sum_{l_2 \in L_2}^{\square}[max_{l_1 \in L_1} sim(l_1, l_2)]}{|L_1| + |L_2|}$$

where: |L_1 |and|L_2 | are the lengths of the token sets for words N1 and N2 respectively.

The output of semantic relations is a coefficient in the range (0, 1), indicating the strength of the name similarity. High values correspond to similar names (i.e. 1 indicates identical names), whereas low values correspond to different names.

The semantic similarity measure between geospatial concepts presented here is based on two sets of analyses. The first examined the semantic similarity of all corresponding feature pairs of OS and OSM data for urban and rural areas. It can be seen from the data in Figure 2 (a and b) that there is agreement between the semantic matching rates in both urban and rural areas. In both sets, the lower similarity scores (0.00 to 0.25) predominate whilst there are few features exhibiting high similarity scores (0.75 to 1.00) – in fact none in the urban area. Thus, one of the more significant findings to emerge from this part of the study is that there is a significant separation between most of the feature pairs of OS and OSM datasets.

`

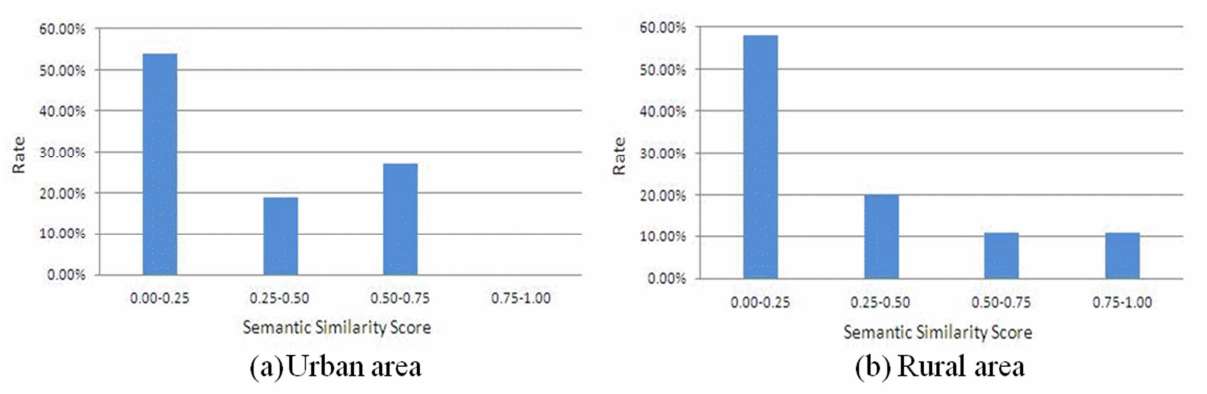

Figure 2. Results of feature based approach

A second set of experiments was implemented to observe the nature of the relationship between the nodes of two schemas. As Howard et al. (2010) reported, there are many categories of nodes relationships ('correspondences') such as One-To-One, when one class in the first schema matches only one class from another schema; One-To-Many, when one node in the source schema matches two or more nodes in the other; and Many-To-One, when multiple instances in one schema match a single instance in another schema. In addition, there are missing correspondences (Source Lacks Data) when the element required by the target schema is lacking in the source schema or Target Lacks Data, when data present in the source schema has no corresponding location in the target schema. A matrix of semantic similarity scores was created for the classes in two schemas (the OS schema and the OSM schema). If the semantic similarity score exceeded 0.5 then a relationship was assumed.

The relationships were examined and Tables 1 and 2 show the results obtained from the analysis of semantic correspondence of schema node classifications in the urban and rural areas respectively. The most striking result is that there are no One-To-One relations for any classes in either area. Thus, there is no semantic similarity score greater than 0.5 recorded for a unique relation between only one class in each of the datasets (the Source and the Target classification schemes). Confused relationships are indicated by the significant rate values for the multiple node relations. For instance, in Table 1 the rate of the relations One-To-Many and Many-To-One were 44% and 53% respectively (note that some Source classes do appear in both of these categories). Approximately half of the nodes in one schema match two or more nodes in the other schema. Further, there are considerable numbers of features which have no correspondences in the other schema, reflected by the results for the categories 'Missing correspondences'. These findings confirm the disappointing rate of similarity between formal sources such as OS data and UGC data such as OSM information, and the probable difficulties in integrating their classification schema.

A major reason for such discrepancy is that the tags (names of classes) of OSM features can be arbitrary strings. There is no definitive list of allowed tags in OSM, and the mapper can (and does) choose whatever tags he/she likes. This open condition has enabled mappers to create and use most tags without needing to refer to a central authority or complex decision making process. Although proposals for new tags should be voted upon, in reality only a very small number of people are interested in this process. Many of the contributors just create the tags that they need for their own mapping without vote. Often they simply consult their peers by e-mail and if a suggested tag seems sensible or if no objection is raised, then the mapper may introduce and use a tag regardless of whether anyone else has used it, or whether it has been put to any type of 'vote'.

Table 1. Results of schemas relationships in Cramlington

Node relations	*Rate*
Single correspondences (One class in Source-To-One class in Target)	0%
Single correspondences (One class in Source-To-Many classes in Target)	44%
Single correspondences (Many classes in Source-To-One class in Target)	53%
Missing correspondence (Source Lacks Data)	47%
Missing correspondence (Target Lacks Data)	41%

Table 2. Results of schemas relationships in Clara Vale

Node relations	*Rate*
Single correspondences (One class in Source-To-One class in Target)	0%
Single correspondences (One class in Source-To-Many classes in Target)	43%
Single correspondences (Many classes in Source-To-One class in Target)	57%
Missing correspondence (Source Lacks Data)	43%
Missing correspondence (Target Lacks Data)	39%

4. CONCLUSION AND OUTLOOK

This study has considered and reported on geometric accuracy and integration of data of variable levels of precision, and further addressed similarity of content and possibilities of matching attribute information.

Semantic interoperability is a very important step for geospatial data integration. The main issues are concerned with the relationship between the elements of a classification scheme (the 'legend', in traditional terminology) and the meanings they carry. This paper presents an approach to measure semantic similarity among data sets for such integration purposes. Several ideas from other fields have been combined together in order to assess the integration of official data, such as Ordnance Survey data, and volunteered geographic information, such as OpenStreetMap data. Firstly, a tokenization was applied, splitting up category names composed of multiple words into single words, and considering word roots and common stems. Then, the feature classifications were modelled as an XML schema labelled rooted tree. The WordNet::Similarity software package was used to compute the semantic similarity between the classes. The results of this analysis found that the semantics data of OSM does not match the OS data sets.

In future work, further parameters such as temporal accuracy, completeness, lineage, and other inherent processing (e.g. generalisation) will be incorporated into the integration model. Additional data sources such as images (e.g. Flickr), textual descriptions and user updating (e.g. TomTom MapShare) are increasingly being used to contribute to 'crowdsourced' data sets and their accuracy also needs to be examined.

REFERENCES

Al-Bakri, M. and Fairbairn, D. (2010) 'Assessing the accuracy of crowdsourced data and its integration with official spatial data sets', The Ninth International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences. University of Leicester / UK, pp. 317-320.

Ali, A. B. H. (2002) 'Moment representation of polygons for the assessment of their shape quality', J Geograph Syst, 4, pp. 209–232.

Austin, R. F. (1984) 'Measuring and representing two dimensional shapes', in Spatial statistics and models. Dordrecht: Reidel D, pp. 293-312.

Banerjee, S. and Pedersen, T. (2003) 'Extended gloss overlaps as a measure of semantic relatedness', Proceedings of the 18th international joint conference on artificial intelligence. Acapulco, Mexico, pp. 805-810.

Butenuth, M., Gösseln, G. v., Tiedge, M., Heipke, C., Lipeck, U. and Sester, M. (2007) 'Integration of heterogeneous geospatial data in a federated database', ISPRS Journal of Photogrammetry and Remote Sensing, 62, (5), pp. 328-346.

Chen, C.-C. (1993) 'Improved moment invariants for shape discrimination', Pattern Recognition, 26, (5), pp. 683-686.

Congalton, R. G. and Green, K. (2009) 'Positional accuracy', in Assessing the accuracy of remotely sensed data:priciples and practices. USA: Taylor & Francis Group, LC, pp. 19-54.

Ebdon, D. (1985) Statistics in geography: A practical approach 2nd ed United Kingdom: Wiley-Blackwell.

Esa, R., Mikko, S. and Janne, H. (2006) 'A new convexity measure based on a probabilistic interpretation of images', IEEE Trans. Pattern Anal. Mach. Intell., 28, (9), pp. 1501-1512.

Goodchild, M. F. and Hunter, G. J. (1997) 'A simple positional accuracy measure for linear features', International Journal of Geographical Information Science, 11, (3), pp. 299-306.

Haklay, M. M. (2010) 'How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets', Environment & Planning B, 37, pp. 682 -703.

Hirst, G. and Onge, D. S. (1998) 'Lexical chains as representation of context for the detection and correction malapropisms', in Fellbaum, C.(ed), WordNet: an electronic lexical database. Cambridge: MIT press, pp. 305-332.

Howard, M., Payne, S. and Sunderland, R. (2010) Technical guidance for the INSPIRE schema transformation network service.

Hu, M.-K. (1962) 'Visual pattern recognition by moment invariants', IRE Transactions on Information, 8, pp. 179-187.

Jiang, J. J. and Conrath, D. W. (1997) 'Semantic similarity based on corpus statistics and lexical taxonomy', International Conference Research on Computational Linguistics (ROCLING X). Taiwan, pp. 19-33.

Leacock, C. and Chodorow, M. (1998) 'Combining local context with WordNet similarity for word sense identification', in Fellbaum, C.(ed), WordNet: A Lexical Reference System and its Application. Cambridge: MIT Press, pp. 265-283.

Lin, D. (1998) 'An information-theoretic definition of similarity', Proceedings of the Fifteenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc., pp. 296-304.

OrdnanceSurvey (2010) Ordnance Survey MasterMap. Available at: http://www.ordnancesurvey.co.uk/osmastermap/ (Accessed: 10 December 2010).

Patwardhan, S., Banerjee, S. and Pedersen, T. (2003) 'Using measures of semantic relatedness for word sense disambiguation', Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics (CICLING-03). Mexico city, pp. 241-257.

Pedersen, T., Patwardhan, S. and Michelizzi, J. (2004) 'WordNet::Similarity - Measuring the Relatedness of Concepts', The Nineteenth National Conference on Artificial Intelligence (AAAI-2004). California, pp. 1024-1025.

Ramm, F., Topf, J. and Chilton, S. (2011) 'Making the Free World Map', in OpenStreetMap - Using and Enhancing the Free Map of the World. Cambridge, England: UIT Cambridge Ltd., pp. 3-8.

Resnik, P. (1995) 'Using information content to evaluate semantic similarity in a taxonomy', Proceedings of the 14th international joint conference on Artificial intelligence - Volume 1. Montreal, Quebec, Canada, Morgan Kaufmann Publishers Inc., pp. 448-453.

Stojmenovi´c, M. and uni´c, J. (2008) 'Measuring Elongation from Shape Boundary', J. Math. Imaging Vis., 30, (1), pp. 73-85.

Tansalarak, N. and Claypool, K. T. (2007) 'QMatch - Using paths to match XML schemas', Data & Knowledge Engineering, 60, (2), pp. 260-282.

Tveite, H. (1999) 'An accuracy assessment method for geographical line data sets based on buffering', International Journal of Geographical Information Science, 13, (1), pp. 27-47.

Uitermark, H. T., Oosterom, P. J. M. v., Mars, N. J. I. and Molenaar, M. (1999) 'Ontology-based geographic data set integration', International Workshop on Spatio-Temporal Database Management STDBM'99. Edinburgh, Scotland, UK, pp. 60-78.

Wu, Z. and Palmer, M. (1994) 'Verb semantics and lexical selection', 32nd. Annual Meeting of the Association for Computational Linguistics. Las Cruces, New Mexico, USA, pp. 133-138.