Conference or Workshop Item (Paper)

Wadsworth, Richard A.; Comber, Alexis J.; Fisher, Peter F. 2008 Probabilistic latent semantic analysis as a potential method for integrating spatial data concepts. In: Navratil, Gerhard, (ed.) *Proceedings of the Colloquium for Andrew U. Frank's 60th Birthday 2008.* Vienna, Department of Geoinformation and Cartography, 99-108. (GeoInfo, 39).

PROBABILISTIC LATENT SEMANTIC ANALYSIS AS A POTENTIAL METHOD FOR INTEGRATING SPATIAL DATA CONCEPTS

R.A. Wadsworth[1], A.J. Comber[2], P.F. Fisher[2],

[1]CEH Lancaster, Bailrigg, Lancaster, LA1 4AP, UK. E-mail: rawad@ceh.ac.uk
[2]Department of Geography, University of Leicester, Leicester, UK. E-mail: ajc36@le.ac.uk, pff1@le.ac.uk

ABSTRACT

In this paper we explore the use of Probabilistic Latent Semantic Analysis (PLSA) as a method for quantifying semantic differences between land cover classes. The results are promising, revealing 'hidden' or not easily discernable data concepts. PLSA provides a 'bottom up' approach to interoperability problems for users in the face of 'top down' solutions provided by formal ontologies. We note the potential for a meta-problem of how to interpret the concepts and the need for further research to reconcile the top-down and bottom-up approaches.

1. INTRODUCTION

Many workers have identified differences in data semantics as *the* major barrier to data integration and interoperability (Frank, 2001; Harvey et al., 1999; Pundt & Bishr, 2002) and as Frank (2007a) notes, "In order to achieve interoperability in GIS, the meaning of data must be expressed in a compatible description". The crux of the problem is that the same real world features can be represented in different ways. The suitability (quality) of a data set is therefore not static or absolute but depends on the appropriateness of the representation in the context of the user's needs (Frank et al., 2004; Frank, 2007b).

Many large datasets depend on multi-disciplinary teams whose members have different conceptualizations of the phenomena being recorded, and who are funded by Research Councils, Government Departments and

Conservation Agencies etc who bring their own set of policy, scientific, financial and ethical concerns to the process. The difficulty in achieving interoperability in this context has not been helped by it becoming enmeshed in narrow technical issues related to discovery metadata and metadata reporting standards.

A "top-down" approach to interoperability might start with the formal assertion that Newtonian physics and Euclidian geometry are sufficient (Frank 2003) and proceed to the development of ontologies, taxonomies and controlled vocabularies into which real data may be placed. We adopt a "bottom-up" approach and consider interoperability from the standpoint of a (naive) data user. We want to know; what the data "labels" mean, how the categories are related to each other and did the data producer have the same conceptual understanding of the phenomenon as the user? The final, and perhaps most difficult, task of bridging the top-down and bottom-up approaches has yet to be attempted within both formal ontology research activities such as OWL and emerging e-science infrastructures such as INSPIRE.

In particular we are concerned with *consistency* and *similarity* between data objects and how this affects a user's analysis (Comber et al., 2006). This paper proposes a text mining approach called Probabilistic Latent Semantic Analysis (PLSA) (Hofmann 1999a,b) to extract or infer the data concepts contained in written descriptions of spatio-environmental information.

2. ESTIMATING SEMANTIC CONSISTENCY.

Estimating Semantic consistency can be done in various ways:

Declarative Approaches: Rules (typically *If ... Then ... Else ...*), are used to characterize relationships between objects. Generating rules is difficult, time consuming, and error prone (especially inconsistent rules).

Semantic Look Up tables: Relationships are encoded in tables (matrices). Comber et al., (2004a,b; 2005a,b) used expert opinion to encode consistency as "expected", "uncertain" and "unexpected" relationships in a successful attempt to compare two Land Cover Maps – a problem the data producers warned users was intractable. Wadsworth et al (2005) decomposed land cover attributed into data primitives before re-integrating them to explore inconsistencies between three land cover

maps of Siberia. Fritz & See (2005) used fuzzy logic to average the response of a group of experts.

Statistical approaches: Foody (2004), Hagen (2003), Csilliag & Boots (2004) used statistical analysis to compare alternative representations of the same phenomenon in attempts to highlight the locations where variables are incompatible. Kampichler et al., (2000), Maier & Dandy (2000), Guo et al., (2005) and Phillips et al., (2006) made use of Genetic Algorithms and Neural Networks for similar purposes. These approaches are not always robust in the face of "noise".

The first two approaches (declarative and semantic) rely on the interaction with domain experts (knowledge engineering). As experts are not always available we try and extract the knowledge they have "stored" in written descriptions. NLP (natural language processing), especially of scientific texts, is a very complex problem but document categorization and information retrieval making the "bag-of-words" assumption is a much simpler problem. We adapted the work of Lin (1997) and Honkela (1997) to look at the similarity between categories rather than documents (Wadsworth et al., 2006). In an attempt to understand why two categories might be considered similar we now investigating the potential of Probabilistic Latent Semantic Analysis (Hofmann 1999a,b).

3. METHODS

In Latent Analysis the assumption is that there are underlying and unobserved variables (the latent variables) that can be used to explain an observed pattern. In Latent Semantic Analysis the pattern is the frequency of words in documents and the latent variables are concepts (ideas) described in the documents. We can observe the relationship between the documents and words and we want to uncover the latent concepts that can explain the distribution of words in documents. Probabilistic Latent Semantic Analysis (PLSA) was proposed by Hofmann (1999a,b) as a "generative" model of latent analysis; the joint probability that a word (w) and document (d) co-occur (P(d,w)) is a function of the conditional probability that the document contains a concept (z) (P(z|d)) and the conditional probability that the word is associated with that concept (P(w|z) (equation 1)

$$P(d,w) = P(d)\sum_{z \in Z} P(w \mid z)P(z \mid d) \qquad \text{Eq. 1}$$

Because we know the frequency of the words in documents it is possible to rearrange the probabilities to develop an iterative expectation maximization scheme to generate P(z|d,w) (the expectation step) and P(w|z), P(d|z) and P(z) (the maximization step). There can be problems with "over-fitting" so Hofmann (1999a) proposes using a variation on simulated annealing (called tempered annealing) to prevent this. Unfortunately the tempered annealing requires a "hold out" of test data and some of our data sets are too short to allow this.

Deciding how many latent variables exist is analogous to determining how many classes exist in a fuzzy classification scheme (like c-means). Making the assumption that the probabilities are like membership functions then the indices proposed by Roubens (1982) can be applied. We use both the Fuzziness Performance Index (FPI) and the Modified Partition Entropy (MPE) and estimate the "best" number of classes from where the sum of the two indices is at a minimum.

Because of restrictions on space we present the results of PLSA for only a single data set, the Land Cover Map of Great Britain (LCMGB; Fuller et al., 1994) class descriptions.

4. RESULTS

4.1 Number of latent variables in the LCMGB

The optimum number of latent variables is about 12; the minimum of the combined FPI and MPE (Roubens 1982). Because the process may converge to a local minima five trials were conducted; Figure 1.

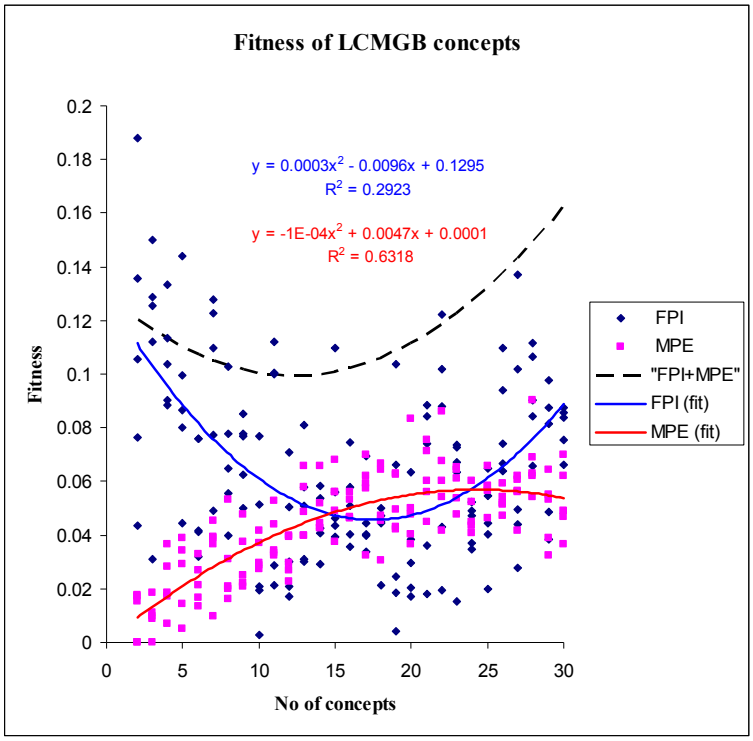

Figure 1. Fitness measures used to determine optimum number of latent variables (concepts) in the LCMGB categories.

4.2 LATENT VARIABLES UNCOVERED BY PLSA

Table 1 shows the relationship between the LCMGB classes and the PLSA concepts.

Table 1.
Twelve Concepts in the LCMGB Descriptions

LCMGB class	Concept	Concept key words (**bold** Prob' > 0.99, *italics* Prob' > 0.5)
0 - unclassified 10 - Open Shrub Moor 16 - Coniferous evergreen woodland (0.54)	J	**25, within, unclassified, types, ten-year, target, rhododendron, regrowth, defined, cycle, coniferous, cloud, classification, cells, cell, burnt, 25metre, 1km**, *both, some, heather, cover*
1 - Sea Estuary 12 - Bracken	F	**Fern, estuaries**, *sea, point, bridging*
2 - Inland water	C	**Fresh**, *estuarine, waters*
3 - coastal bare 4 - Saltmarsh	G	**Tides, tide, spring, sometimes, silt, seaweed, saltmarsh, rocks, prevailed, normal, mud, intertidal, habitats, discrepancies, dependent, define, days, beds, arise,** *shingle, lower, saltmarshes, beach, limit, imaging, vegetated, sand, level, green, grasses, cliffs, high*

5 - Grass Heath 9 - Moorland Grass 17 - Upland bog 24 - Lowland bog	K	**Typically, temporary, scotland, nardus, myrica, molinia, logging, eriophorum, deschampsia, bogs, bog, ammophila, agricultural,** *hill, dunes, acid, west, north, wetlands, unenclosed, might, upland, heaths, lowland*
6 - Mown Grazed Turf	L	**Material,** *throughout, sward, amenity, turf, mown, use, uncropped, grazed*
7 - Meadow Verge Semi-natural swards	H	**Verges, transition, low, intensity, improvement, hay, appearance,** *agrostis, meadows, cropped, pastures, semi-natural*
8 - Rough Marsh Grass 23 - Felled Forest	I	**Forest, felled,** *rough, marsh*
11 - Dense Shrub Moor 13 - Dense Shrub Heath 15 - Deciduous Woodland 25 - Open Shrub Heath 16 - Coniferous evergreen woodland (0.46)	E	**Unique, separates, ling, heathland, evergreen, erica, area,** *dense, woodland, mixed, heath, shrub, however, burning, mixture, deciduous*
14 - Scrub Orchard 18 - tilled land (arable crops) 20 - Suburban rural development (0.75)	B	**Tilled, temporarily, suburban, sallow, rural, roads, orchards, orchard, hawthorn, few, buildings,** *small, scrub, trees, land, pixels*
19 - Ruderal weed	A	**Aside,** *weed, set, ruderal*
21 - Urban development 22 - Inland bare ground	D	**Urban, surfaces, developments, car,** *associated, fill, industrial, parks, permanent, gravel, major, sites, development*

5. DISCUSSION AND CONCLUSIONS

Although the description of the LCMGB classes are rather short the PLSA has managed to identify some reasonable concepts, (reasonable in the eyes of a domain expert). Unfortunately, some of the concepts are rather more difficult to interpret and may reflect statistical artifacts or the lack of words to process. When applying the approach to other data sets (not reported here) we have had mixed results, for soils the pattern is very complex and not readily interpretable. On the other hand applying it to the abstracts from the IJGIS produced readily interpretable "clusters".

Where human domain experts exist then knowledge engineering methods can codify their expertise in ways that make inter-operability a practical proposition. Domain experts may not exist or may not be accessible (through time constraints or geography) in those cases where domain experts have expressed their expertise through *long* textual descriptions simple text mining can produce acceptable estimates of semantic similarity. A "reconnaissance" assessment of PLSA suggests

that it may go someway to explain why concepts are considered to the similar. As yet the task of reconciling the top-down and bottom-up approaches to interoperability remain unexplored but the PLSA approach can be applied to more than one dataset to identify classes (ie documents) with shared concepts to facilitate data integration.

REFERENCES

Comber, A., Fisher, P., Wadsworth, R., (2004a). Integrating land cover data with different ontologies: identifying change from inconsistency. International Journal of Geographical Information Science, 18(7): 691-708.

Comber A.J.. Fisher P.F., Harvey, F., Gahegan., M and Wadsworth R.A, (2006). Using metadata to link uncertainty and data quality assessments. pp 279–292 in Progress in Spatial Data Handling, Proceedings of SDH 2006, (eds. Andreas Riedl, Wolfgang Kainz, Gregory Elmes), Springer Berlin.

Comber, A.J., Fisher, P.F., Wadsworth, R.A., (2004b). Assessment of a Semantic Statistical Approach to Detecting Land Cover Change Using Inconsistent Data Sets. Photogrammetric Engineering and Remote Sensing, 70(8): 931-938.

Comber, A.J., Fisher, P.F., Wadsworth, R.A., (2005a). A comparison of statistical and expert approaches to data integration. Journal of Environmental Management, 77: 47-55.

Comber, A.J., Fisher, P.F., Wadsworth, R.A., (2005b). Combining expert relations of how land cover ontologies relate. International Journal of Applied Earth Observation and Geoinformation, 7(3): 163-182.

Csillag, F. and Boots, B., (2004), Toward comparing maps as spatial processes, pp. 641 – 652, Developments in Spatial Data Handling, P. Fisher (Ed.), Springer Verlag, Berlin

Foody G.M., (2004). Thematic map comparison: Evaluating the statistical significance of differences in classification accuracy, Photogrammetric Engineering and Remote Sensing, 70 (5): 627-633.

Frank, A.U., (2001). Tiers of ontology and consistency constraints in geographical information systems, International Journal of Geographical Information Science, 15 (7): 667-678.

Frank, A.U. (2003) A linguistically justified proposal for a spatio-temporal ontology. **COSIT'03,** Conference on Spatial Information Theory 24-28 September, Ittingen, Switzerland

Frank AU, Grum E, Vasseur R (2004). Procedure to select the best dataset for a task, GEOGRAPHIC INFORMATION SCIENCE, LECTURE NOTES IN COMPUTER SCIENCE Volume:3234:81-93

Frank A.U., (2007a). Towards a Mathematical Theory for Snapshot and Temporal Formal Ontologies Pages 317-334 in The European Information Society, Lecture Notes in Geoinformation and Cartography, Springer Berlin Heidelberg

Frank A.U., (2007b).Incompleteness, error, approximation, and uncertainty: An ontological approach to data quality pp 107-131 in GEOGRAPHIC UNCERTAINTY IN ENVIRONMENTAL SECURITY (eds Morris A, Kokhan S), Proceedings of NATO Advanced Research Workshop on Fuzziness and Uncertainty in GIS for Environvmental Security and Protection Kyiv, UKRAINE, JUN 28-JUL 01, 2006

Fritz, S. and See, L., (2005). Comparison of land cover maps using fuzzy agreement. International Journal of Geographical Information Science 19(7) 787-807.

Fuller, R. M., Groom, G. B., and Jones, A. R., 1994, The Land Cover Map of Great Britain: an automated classification of Landsat Thematic Mapper data. Photogrammetric Engineering and Remote Sensing, 60, 553–562.

Guo, Q.H., Kelly, M., Graham, C.H., (2005). Support vector machines for predicting distribution of sudden oak death in California, Ecological Modelling 182 (1): 75-90.

Hagen, A., (2003), Fuzzy set approach to assessing similarity of categorical maps, International Journal of Geographical Information Science , 17(3): 235-249.

Harvey F, Kuhn W, Pundt H, Bishr Y, Riedemann C (1999) Semantic interoperability: A central issue for sharing geographic information, Annals of Regional Science 33(2): 213-232

Hofmann, T (1999a). Probabilistic latent semantic indexing, pp 50-57 in Proceedings of 22nd International Conference on Research and Development in Information Retrieval (Eds Hearst M, Gey F, Tong R) Univ Ca, Berkeley, California, Aug, 1999

Hofmann, T (1999b). Probabilistic latent semantic analysis, pp 289-296 in Proceedings of 15th Conference on Uncertainty in Artificial

Intelligence (Eds. Laskey KB, Prade H) Royal Inst Technol, Stockholm, Sweden, Jul 30-Aug 01, 1999

Honkela T., (1997). Self-Organising maps in natural language processing. PhD thesis Helsinki University of Technology, Department of Computer Science and Engineering, http://www.cis.hut.fi/~tho/thesis/

Kampichler, C., Dzeroski, S. and Wieland, R., (2000). The application of machine learning techniques to the analysis of soil ecological data bases: relationships between habitat features and Collembola community characteristics. Soil Biology and Biochemistry 32: 197–209.

Lin, X., (1997). Map displays for information retrieval. Journal of the American Society for Information Science, 48:40-54.

Maier, H.R. and Dandy, G.C., (2000) Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications, Environmental Modelling and Software, 15, 101-124,

Phillips, S.J., Anderson, R.P., Schapire, R.E., (2006). Maximum entropy modeling of species geographic distributions, Ecological Modelling 190 (3-4): 231-259.

Pundt, H. and Y. Bishr, (2002). Domain ontologies for data sharing-an example from environmental monitoring using field GIS. Computers and Geosciences, 28 (1): 95-102.

Roubens, M., (1982). Fuzzy clustering algorithms and their cluster validity. Eur. J. Oper. Res. 10, 294–301.

Wadsworth R.A, Comber A.J., and Fisher P.F., (2006). Expert knowledge and embedded knowledge: or why long rambling class descriptions are useful. pp 197 – 213 in Progress in Spatial Data Handling, Proceedings of SDH 2006, (eds. Andreas Riedl, Wolfgang Kainz, Gregory Elmes), Springer Berlin.

Wadsworth R.A., Fisher P.F., Comber A., George C., Gerard F. & Baltzer H. (2005). Use of Quantified Conceptual Overlaps to Reconcile Inconsistent Data Sets. Session 13 Conceptual and cognitive representation. Proceedings of GIS Planet 2005, Estoril Portugal 30th May - 2nd June 2005. ISBN 972-97367-5-8. 13pp.