COSIT at Twenty: Measuring Research Trends and Interdisciplinarity

Karl Grossner[1] and Benjamin Adams[2]

University of California, Santa Barbara
[1]Department of Geography; [2]Department of Computer Science
[1]karlg@geog.ucsb.edu; [2]badams@cs.ucsb.edu

Abstract. We have performed a topic classification procedure on a text corpus consisting of the proceedings for all ten meetings of the Conference on Spatial Information Theory (COSIT) held between 1992 and 2009, providing a measure with which to answer several kinds of questions about the dynamic conceptual content of that conference series. We have identified topics trending upward and downward, looking particularly at the level of research interest in cultural factors. We have also investigated whether there has been growing interdisciplinarity in the research reported at COSIT meetings by this diverse and dynamic scholarly community. Preliminary results are presented, planned future work is discussed, and additional questions are invited.

1 Introduction

In the twenty years since “COSIT Zero,” the formative 1992 meeting held in Pisa, Italy, a multidisciplinary community of scholars with inter-related research interests concerning the many aspects of spatial information theory has clearly solidified to become a distinctive scholarly domain. Spatial information theory concerns a wide range of scientific issues, which have been characterized as an ordered process [1]: (i) the perception of spatiotemporal phenomena; (ii) cognitive tasks including the creation and manipulation of internal representations like concepts and cognitive maps; (iii) the creation of external representations such as formal-logical systems and maps; (iv) their implementation for advanced reasoning with computing systems; (v) the use and usability of such systems; and (vi) the communication of spatiotemporal knowledge by many means.

In the published proceedings for early COSIT meetings, editors outlined some broad goals for the conference series. It was hoped that focused interdisciplinary dialog would lead to discovery of “universally valid principles” providing a “consistent basis for GIS” [2]. In 2001, Dan Montello wrote that researchers from “(several) specializations within geography, computer science, and psychology” were “increasingly sharing methods and concepts” [3]. The present milestones—ten meetings, twenty years—seem an appropriate time to look closely at the language of all papers and plenary abstracts published in COSIT proceedings, in order to identify trends in research topics and the nature and extent of interdisciplinarity. An additional question, put to us by David Mark, is whether there has been a diminishing focus on the cultural factors pertaining to spatial reasoning. We attempt to answer these questions using natural language processing methods for topic modeling and similarity assessment, described below. Following that we describe and illustrate some preliminary measures and results, then conclude with brief discussion of additional questions to be addressed in the coming months, results from which will be presented in our poster in September 2011.

2 Introduction to LDA

Latent Dirichlet Allocation (LDA) is a probabilistic topic model that describes each document in a text corpus as a unique mixture of latent topics [4]. Each topic in turn is described as a probability distribution over words. LDA makes the assumption that document generation can be explained in terms of these distributions, which are assumed to have a Dirichlet prior. First, a topic distribution is chosen for the document, and then each word in the document is generated by 1) randomly selecting a topic from the topic distribution and 2) randomly selecting a word from the chosen topic. Given a set of documents, the main challenge is to infer the word distributions and topic mixtures that best explain the observed data. This inference is computationally intractable, but an approximate answer can be found using a Gibbs sampling approach [5].

Although this generative model is a simplification of the actual writing process, it provides an unsupervised statistical method for identifying latent topics in text and exploring how documents are similar or different. The semantic distance between documents and document sets can be measured using the symmetric Kullback-Liebler divergence (relative entropy) of their topic distributions:

$$D_{KL} = \sum_i \left[x_i \log \left(\frac{x_i}{\sum_j x_j} / \frac{y_i}{\sum_j y_j} \right) / \sum_j x_j \right]$$

3 Data preparation and LDA topics

The ten conference proceedings published to date contain 308 articles (294 full papers and 14 keynote and poster abstracts). The text of these was extracted from PDF files, stripped of “References” sections, punctuation, numbers, and common connector words (e.g. articles and prepositions). The resulting simple word vectors were stemmed to reduce the variation in word forms (e.g. removing suffixes and plural endings). Several articles could not be processed due to technical issues, and we were left with 296 documents. These were analyzed using the LDA algorithm discussed above to produce 20 topics. For each topic, LDA ranks each term in the corpus, and we have used the top ten terms for each, as illustrated in Figure 1.

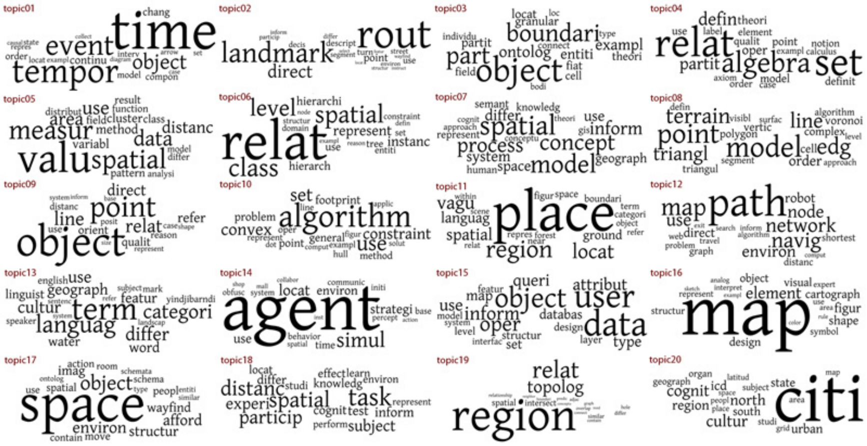

topic01
time tempor event object chang order locat exampl continu interv arrow set model compon collect repres diagram case
topic02
rout landmark direct decis descript turn point street use segment environ particip inform differ
topic03
boundari object part ontolog entiti exampl theori locat granular loc individu partit field fiat cell bodi connect type
topic04
relat algebra set defin theori use label element qualit point notion oper exampl calculus partit axiom order model case definit
topic05
valu measur spatial data distanc area use result function distribut field cluster class method variabl model differ pattern analysi
topic06
relat level hierarchi spatial represent class hierarch instanc tree entiti constraint defin set node structur domain exampl use reason
topic07
spatial concept model process inform system semant knowledg differ cognit approach represent conceptu theori use gis geograph human space
topic08
point model terrain line triangl edg defin visibl surfac vertic polygon algorithm voronoi complex level cell order approach segment triangul
topic09
object point direct line relat refer distanc system inform base posit use orient case shape reason qualit size represent
topic10
algorithm set footprint applic problem convex oper constraint represent general figur dot point comput exampl use solut hull method line
topic11
place region locat vagu within languag scene spatial relat figur space boundari term categori object refer ground repres forest near
topic12
path map use network navig robot node direct exit search inform algorithm web travel problem graph environ distanc shortest comput
topic13
term languag categori geograph use english linguist cultur speaker subject mark featur yindjibarndi refer sentenc system landscap differ water word
topic14
agent locat environ simul strategi use behavior spatial time shop mall obfusc system collabor communic initi base percept action
topic15
data user object inform queri attribut featur map use model system level oper interfac structur set databas design layer type
topic16
map element cartograph object visual expert analog interpret sketch represent exampl structur use area figur rule shape color symbol design
topic17
space object imag action room schemata schema ontolog use spatial type peopl entiti similar wayfind afford environ structur contain move
topic18
spatial distanc task particip experi locat differ studi effect learn knowledg environ cognit test inform represent perform subject
topic19
region relat topolog spatial intersect relationship similar contain hole differ overlap graph
topic20
citi cognit region north south cultur urban organ geograph latitud icd space subject state peopl place area studi grid map

Fig. 1. Twenty COSIT topics identified with LDA analysis

Many of the topics seem coherent; others are less so. Topic coherence, while intuitively satisfying, is not critical for their utility in similarity analyses. Effectively, these topics constitute 20 linguistic conceptual dimensions which are ‘present’ to greater or lesser degree in each document in the corpus. The result is that each document has a 20-value array as a topic signature. This attribute can be joined with others of interest: year of publication, authors’ disciplinary affiliations, and a thematic category from the proceedings. The relative ‘strength’ of topics and topic sets can be traced over time, across all papers or within disciplines. Individual document signatures and their aggregations by year, discipline and author can be easily assembled in matrices and compared for similarity.

4 Question 1: What topics are trending up or down?

We created a topic distribution for each conference year from the LDA results by finding the mean strength of each topic for all papers written in a given year. We then created a 2D map of conference years using a pair-wise matrix of year-by-year semantic distances (Figure 2) which indicates meetings are by-and-large distinctive. Then we performed linear regression (OLS) on each topic over time and ordered the results by slope.

The slopes of topic trends indicate several cases of shifting research focus (Figure 3). The results seem to support the notion that traditional static data representation questions for GIS (as manifested in topics 14, 8, and 7) have been supplanted by research related to navigation and mobile GISs (topics 1 and 11). While this result may not surprise those familiar with the trends in the spatial information community, it does provides additional intuitive validation that the topics extracted by LDA have interpretable meaning.

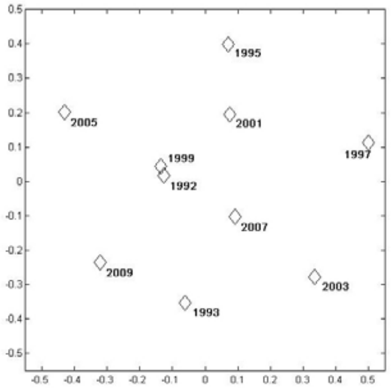
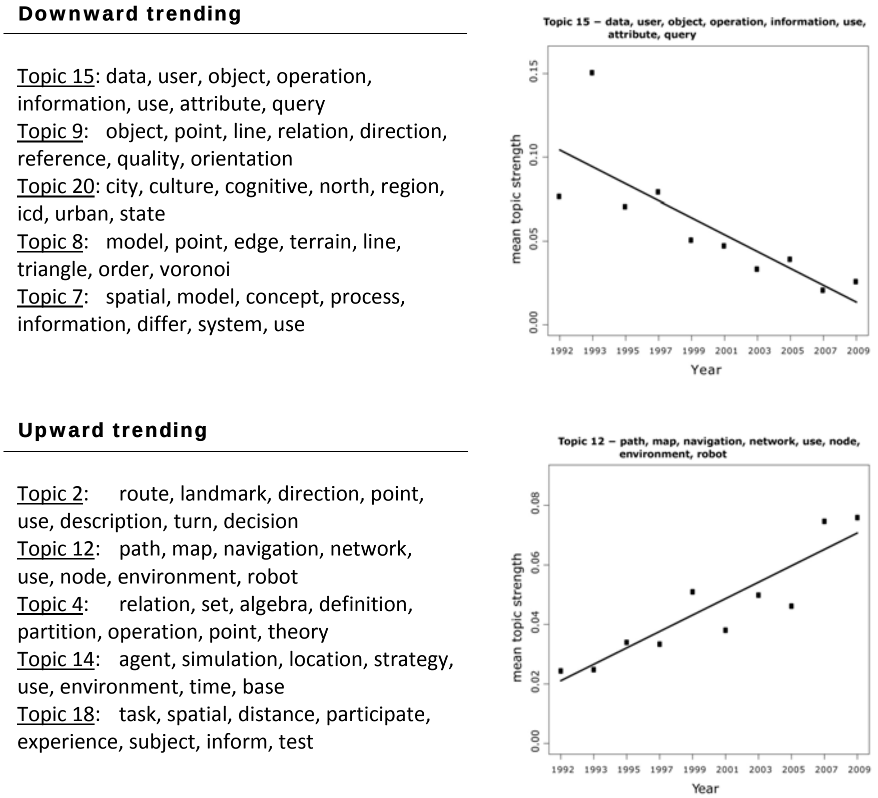


Fig. 2. Year-by-year semantic distances



Downward trending

Topic 15: data, user, object, operation, information, use, attribute, query
Topic 9: object, point, line, relation, direction, reference, quality, orientation
Topic 20: city, culture, cognitive, north, region, icd, urban, state
Topic 8: model, point, edge, terrain, line, triangle, order, voronoi
Topic 7: spatial, model, concept, process, information, differ, system, use

Upward trending

Topic 2: route, landmark, direction, point, use, description, turn, decision
Topic 12: path, map, navigation, network, use, node, environment, robot
Topic 4: relation, set, algebra, definition, partition, operation, point, theory
Topic 14: agent, simulation, location, strategy, use, environment, time, base
Topic 18: task, spatial, distance, participate, experience, subject, inform, test

Fig. 3. Downward and upward trending topics

5 Question 2: What about people in spatial information theory?

We identified topics 13, 20, and 11 as having an explicitly cultural theme, and graphed their prominence year by year (Figure 4). We then selected the 11 papers in Proceedings sections having labels denoting explicitly cultural themes, statistically identified topics 13, 20, 7, and 18 as most associated with those papers and tracked those topics over time (Figure 5). Certainly 2009 saw a drop in focus for two of the 'cultural topics;' there has been an upward trend for 7 and 18, neither of which seem obviously 'cultural.'

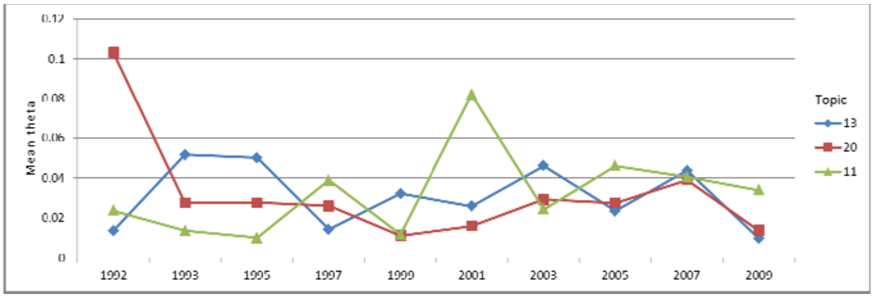


Fig. 4. Explicitly cultural topics, as judged by this paper's authors

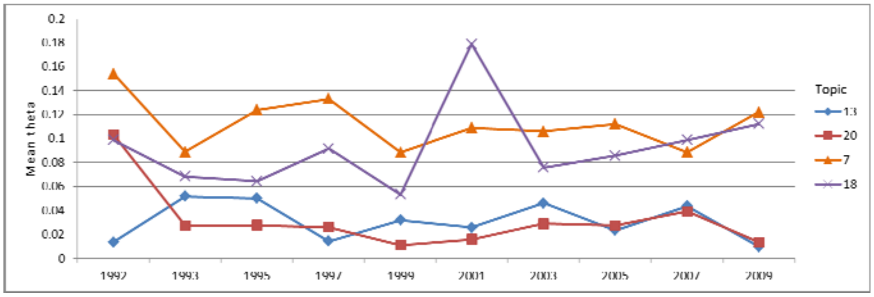


Fig. 5. Explicitly cultural topics, according to Proceedings classifications

6 Question 3: Has interdisciplinary synergy increased?

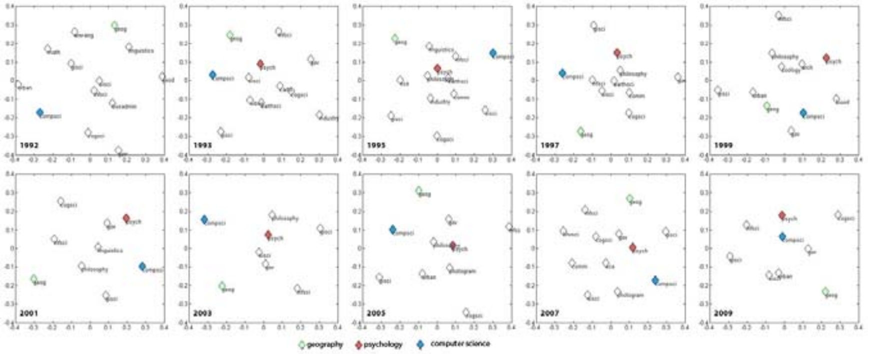


Fig. 6. Semantic distances between disciplines over time

We created disciplinary ‘average topic signatures’ for each year, then calculated a dissimilarity (distance) matrix for each year and an associated 2D plot (Figure 6). As is usual for MDS, the results can be interpreted variously. We can see shifting distances between disciplines, but trends are not conclusive. Nor can we confirm overall increasing interdisciplinarity with this method.

7 Future work

We plan to obtain the text from the 12 missing papers and re-run the operations described above. We will also replicate the process used for Question 2 to look at other broad conceptual categories and further analyze disciplinary ‘movement’ in conceptual space, hoping to identify latent trends. Also, by developing topic-paper graph visualizations, we are expecting to better evaluate trends and disciplinary cross-fertilization.

References

1. Winter, S., Duckham, M., Kulik, L. and Kuipers, B. (2007). Preface. In Winter, S., et al (Eds.): *COSIT 2007, LNCS 4736*, pp.v-vi.
2. Frank, A.U. and Campari, I. (1993). Preface. In A. U. Frank and I. Campari (Eds.): *COSIT 1993, LNCS 716*, p.v.
3. Montello, D. R. (2001). Preface. In D. R. Montello (Ed.): *COSIT 2001, LNCS 2205*, pp.v-vi.
4. Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *Journal of Machine Learning Research* 3 (2003), 993-1022.
5. Griffiths, T. L., and Steyvers, M. Finding scientific topics. *Proceedings of the National Academy of Sciences* 101, Suppl. 1 (April 2004), 5228-5235.