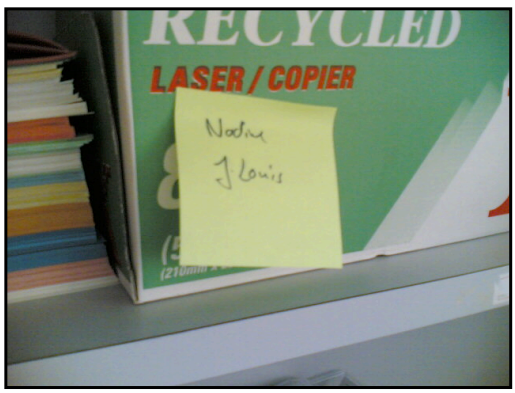
Mobile Search on Ubiquitous Collaborative Annotations of Space

Workshop: The Future of Web Search
Barcelona - May 19-20, 2006

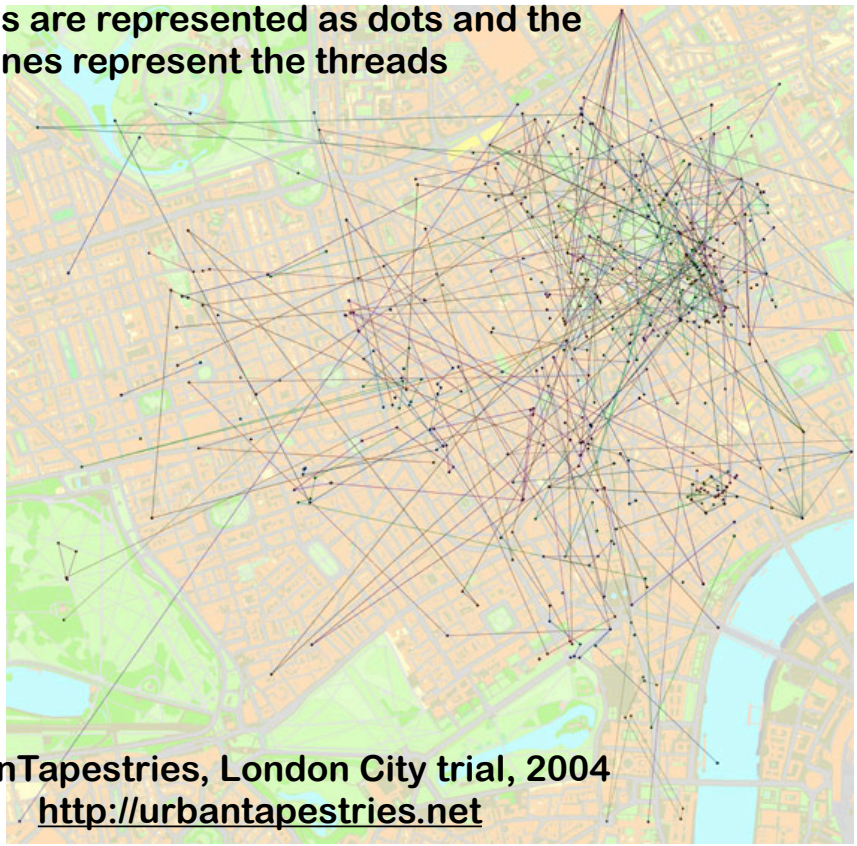
Mauro Cherubini, **Pierre Dillenbourg** (EPFL, CH)
& **Lorenzo Viscanti** (Clipperz, IT)

Spatialised Communication

We use this term to refer to any communication act which makes use explicitly of the geographical context to convey the message.



Messages are represented as dots and the lines represent the threads



UrbanTapestries, London City trial, 2004
http://urbantapestries.net

[21] Lane, 2003

The objective

Supporting mobile information retrieval of ubiquitous and collaborative annotations of space.

Challenges

- Few number of words (5/15)
- Abbreviations (i.e., CU for See you)
- Neologisms & Symbolical language (i.e., 2morrow; :-))
- “community” or vernacular language
- Noise
- synonimy and polisemy

subject	content
First try	Hello world!
have to go	I have to go here, take the TSOL
Computer Supported Collaborative Work	This is the classroom for the course BC03
Current activity	Pierre is giving his course here and not in Odyssea as last year
Re: Computer Supported Collaborative Work	It's a nice first course!
nobody use this tool?!	Why I am the only person using this application?
Impro	Wednesday and Thursday The PIP (Pool d'impro du Poly) have some theatrical imprvisation matches at the Salle Polyvalente
Miam	It's the Kebab week at the Corbusier !
Car	Where's my car ?
Re: nobody use this tool?!	You're not alone anymore ! :-)
Map	The map is too old, the BC building is not on... Is Pierre giving his course in a hut ???
LC	Maybe one day, the Learning Center will stand here...
Pool	And what about a swimming pool here ?
Re: LC	I heard the financial aspects were going quite well, so let's keep faith into the future :)
FIFO day	The FIFO day will be held in collaboration with the Poseidon day in the Hall SG on november 30th.
Chinese Course	12.15 - 14.00 Come too! I'm sure it isn't full yet!
Organizing FIFO day	I have to meet Maureen this week to talk about room reservations and layout of posters during the FIFO day
Restaurant	Beautiful view and terrasse on the last floor of the building. The food is ok, nothing spectacular, but the fresh air compensates :)

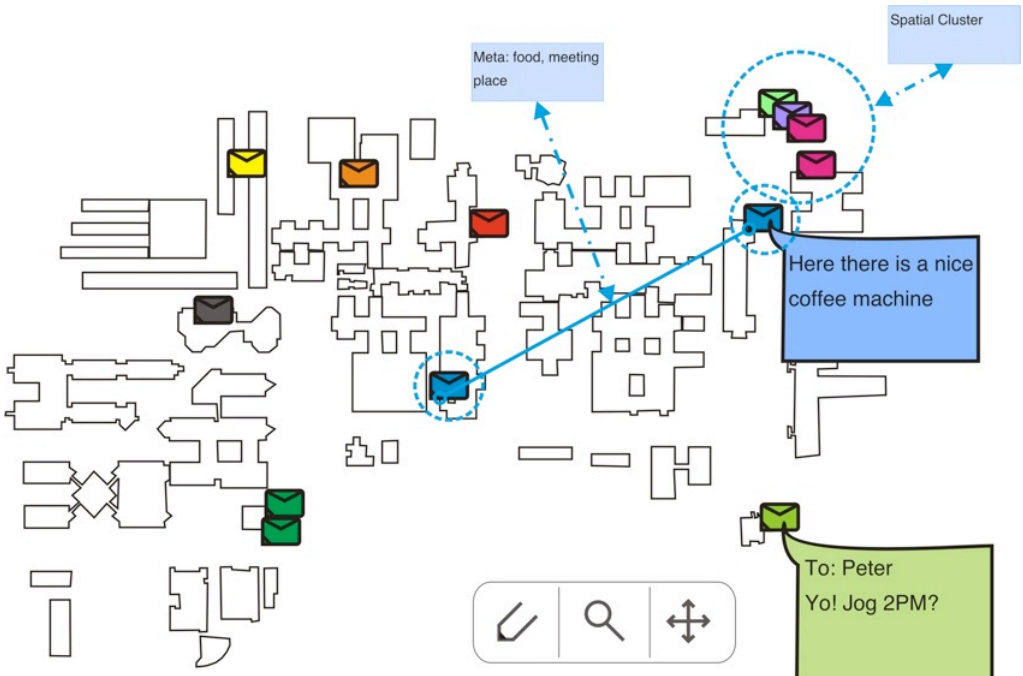
Place codes Location pointers

Supporting what?

- Information retrieval: pointing to the right information that best matches the user needs (finding resources on the city space)
- Information navigation: helping the user to discover what is around

How to support?

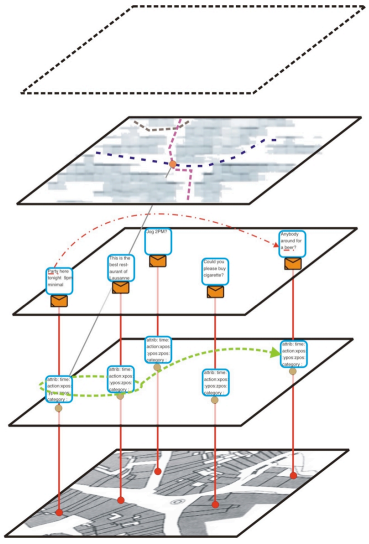
- We need to create links between the messages that presents similarities
- We need to move from single messages to higher level of description and grouping of the information (CLUSTERS)



We are looking for two kinds of relations: links and clusters among the messages

Smart Layering of information

- Messages refer to space and viceversa
- Messages can be treated as pure **geometrical objects**
- or as pure **Semantical objects**



Using these layers we plan to find patterns among the messages

How to create links?

- We can try to combine the geometric and the semantic dimension but how to calculate the semantic similarity?
- LSA (Latent Semantic Analysis) seems one of the most common answers!

Reasons against LSA

- Patented method
- Computationally consuming
- Poor scalability of SVD (Scalable Vector Decomposition)
- Difficult to interpret the underlying reduced term space
- Difficult to select the optimal value of singular values
- expensive pre-processing

An alternative: Contextual Network Graphs

An alternative interpretation of the term-document matrix (TDM): each non-zero value in the TDM correspond to an edge connecting a term node to a document node.

+ CNG does not require pre-processing.

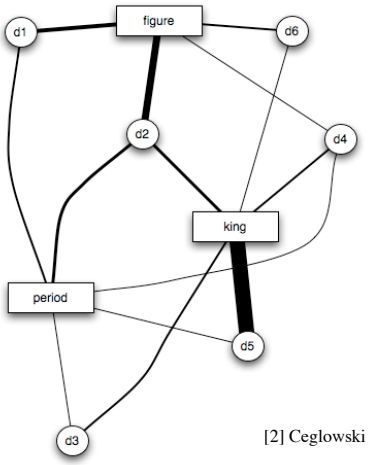
The idea behind

- this representation correspond to the intuition that documents that share many rare terms are likely to be semantically related
- on the same fashion of LSA a query on a particular keyword may still reach a document that does not contain the word itself

Graph-theoretic model

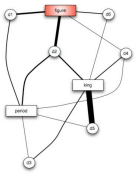
The term-document matrix is interpreted as a connected bipartite graph. Edge weights are based on measures of tf*idf

	d1	d2	d3	d4	d5	d6
figure	4	7	0	1	0	2
king	0	3	2	2	23	1
period	2	3	1	1	1	0



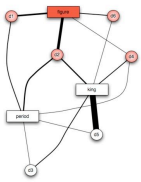
[2] Ceglowski et. al., 2003

Spreading activation



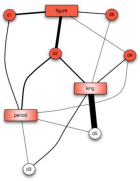
1. One or more nodes are activated

Spreading activation



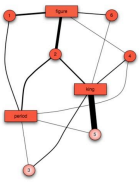
1. One or more nodes are activated
2. Neighbor nodes are activated

Spreading activation



1. One or more nodes are activated
2. Neighbor nodes are activated
3. Energy continues to be distributed through the graph

Spreading activation



1. One or more nodes are activated
2. Neighbor nodes are activated
3. Energy continues to be distributed through the graph
4. Non-keyword results are identified

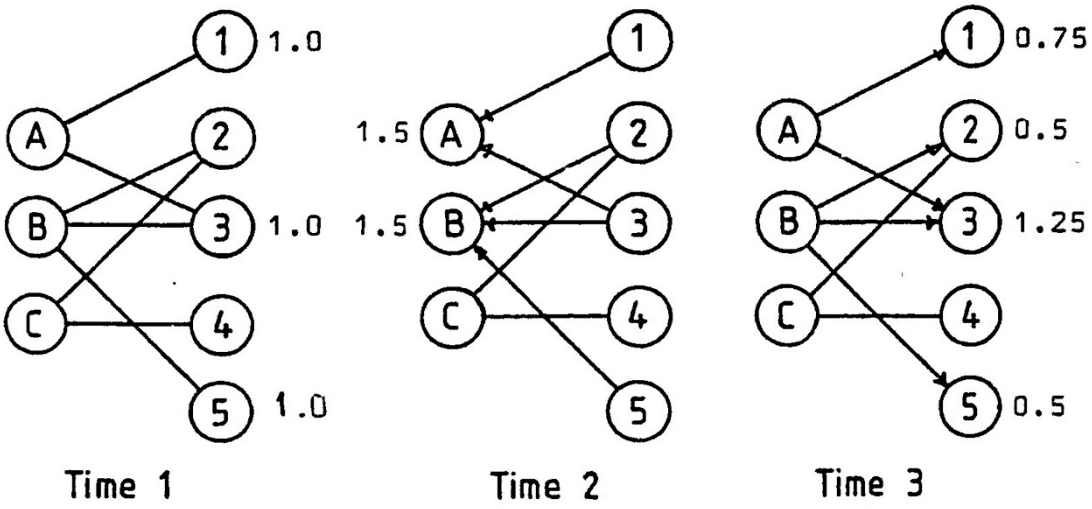
Spreading activation

A searched keyword will energize a node of the network graph and allowing the energy to propagate to other nodes

```
1  procedure energize( energy E, node n_k ) {
2       energy(n_k) := energy(n_k) + E
3       E' := E / degree of  n_k
4       if ( E' > T ) {
5            for each node n_j in N_k  {
6                 E" := E' * e_jk
7                 energize( E", n_j )
8            }
9       }
10 }
```

[2] Ceglowski et. al., 2003

Spreading distribution of energy



[16] Preece, 1981

What is missing?

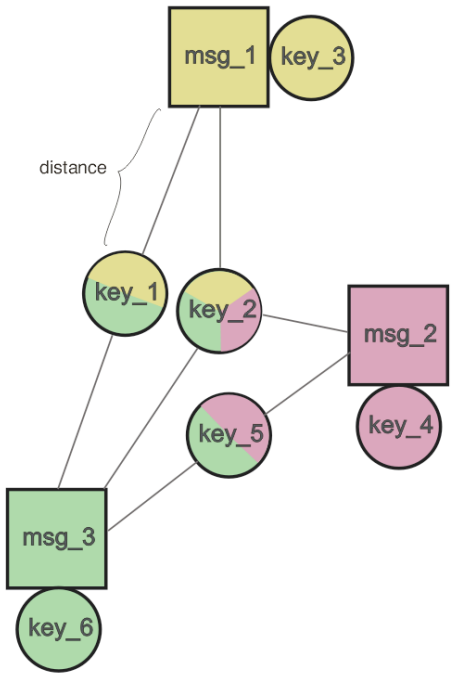
- CNG may be <u>not enough</u> for our particular situation because we'll have tf*idf ≈ 1
- the actual <u>model's topology</u> is arbitrary and is not used to define the spreading schema

Our contribution

- Why not using the geometrical topology of the messages as an energy routing strategy to be used in the spreading activation?

Geographical spreading

- Activation energy can be parametrized using the distance from the message to the keywords given by the geometric distance between the documents (we call this CNG*)

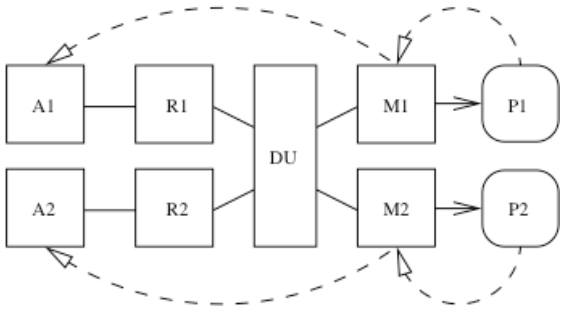


Experimental steps

- phase 1: are LSI and CNG comparable when working on a standard dataset? --complete--
- phase 2: are LSI and CNG comparable when working on a 'spatialized' dataset?
- phase 3: is CNG* better than CNG when supporting information retrieval on a 'spatialized' dataset?

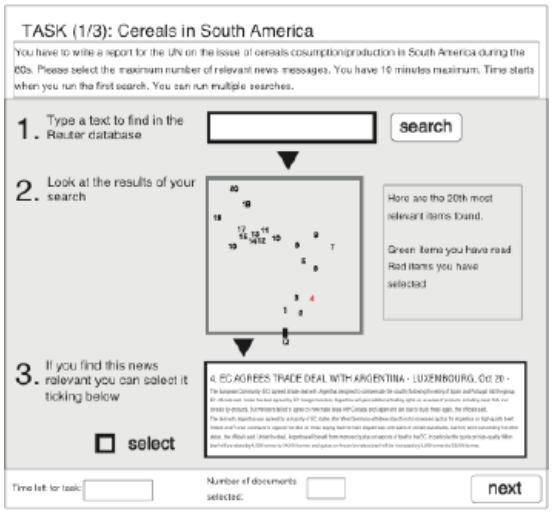
Experimental setup

- Two algorithms (LSA and CNG) produce two ranking lists
- these are processed by a common drawing unit which produces two maps
- using these maps the user achieve certain performances which are referred back to the algorithms



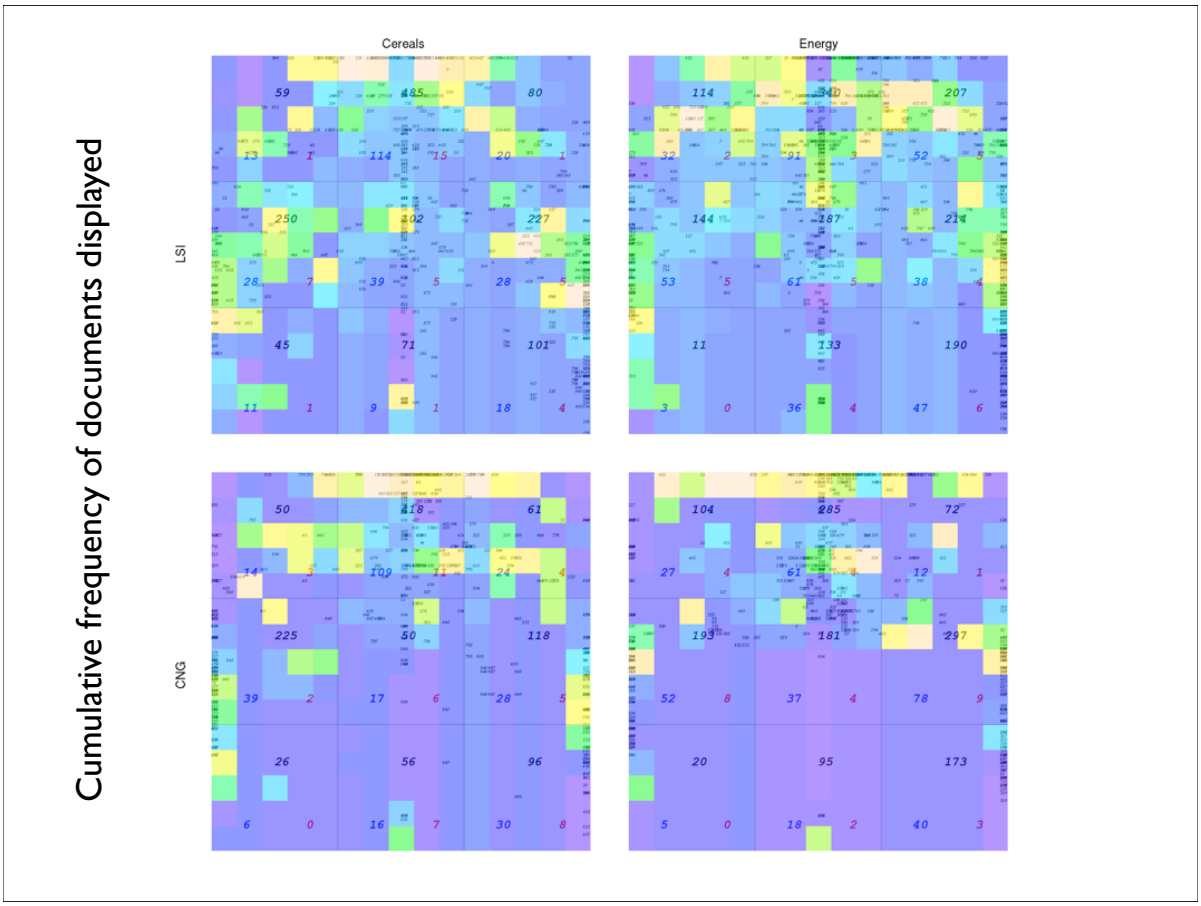
The interface used

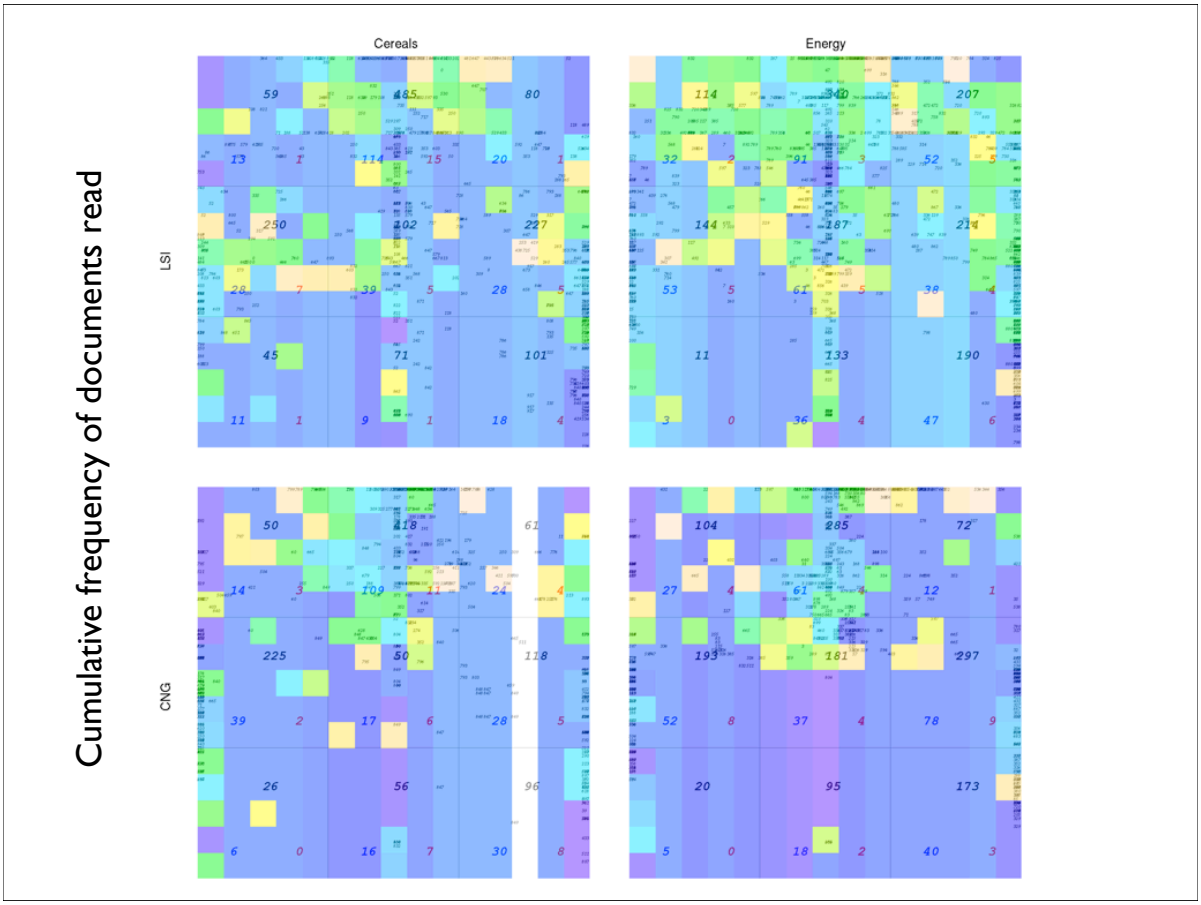
- The user could run multiple queries and see the results in a bi-dimensional map
- The user had to select the maximum number of matching articles in the given time (5 min)

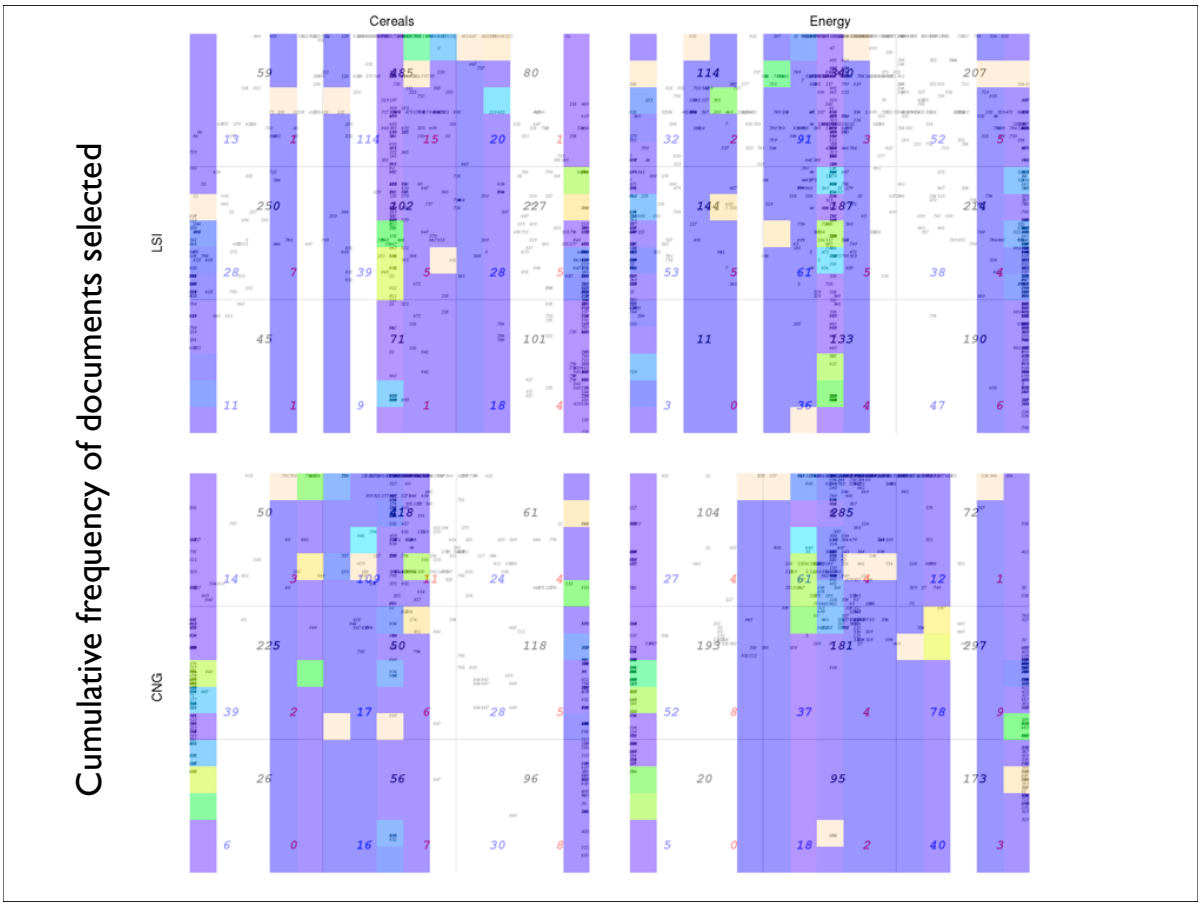


Experimental group

- 47 users passed the experiment
- mainly university students familiar with search systems
- english was not the mother tongue of the majority of the users


Cereals
Energy
LSI
CNG
Cumulative frequency of documents displayed


Cereals
Energy
LSI
CNG
Cumulative frequency of documents read



Some general results

- No significant difference in performances when using LSA compared to CNG
- No significant differences in number of queries or items read or items selected
- No significant differences in time required to select the results

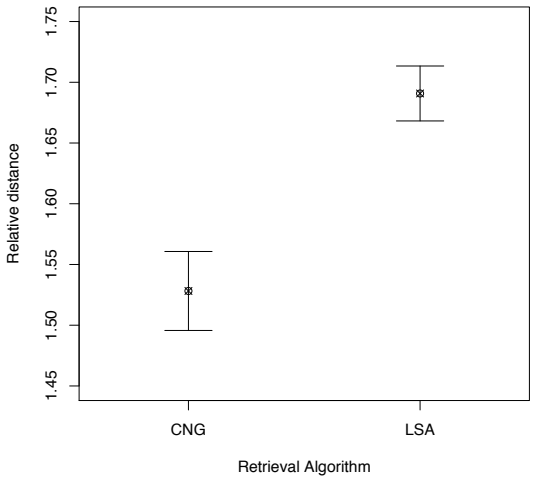
▸ Cheap conclusion: they are comparable

How to define an exploration strategy?

- There is a strategy of exploration <u>if the length of the jump</u> between a document and the following <u>is a function of the starting document pertinence to the task</u>

Was there such strategy?

- Still an open question
- average relative jumps over CNG were lower than those over LSA



Conclusions

- We propose to use a modified spreading activation method CNG* for retrieval of collaborative ubiquitous annotations of space
- To test the validity of this method we propose a 3-phased framework
- The first phase shows that LSA and CNG are comparable when working on a standard dataset

Some missing points

1. We can assign functional meaning to specific keywords like Here; This; ...
2. Is it possible to use CNG to cluster?
3. How can we represent the result on the map to maximise the user experience?

End - Q&A

Thanks!

Mauro Cherubini &
Pierre Dillenbourg (CRAFT - EPFL)
mauro.cherubini@epfl.ch
http://www.i-cherubini.it/mauro/blog/

Lorenzo Viscanti (Clipperz)
lorenzo.viscanti@gmail.com
http://www.noosfactory.com

Bibliography

[1] J. R. Anderson. *The Architecture of Cognition*. Harvard University Press, 1983.

[2] M. Ceglowski, A. Coburn, and J. Cuadrado. Semantic search of unstructured data using contextual network graphs. Preliminary white paper, National Institute for Technology and Liberal Education, Middlebury College, Middlebury, Vermont, 05753 USA, 2003.

[3] F. Crestani. Application of spreading activation techniques in information retrieval. *Artificial Intelligence Review*, 11(6):453–482, December 1997.

[4] S. Deerwester, S. T. Doumas, T. K. Landauer, G. W. Furnas, and R. Harshman. Indexing by latent semantic analysis. *Journal of the Society for Information Science*, 41(6):391–407, 1990.

[5] M. Hearst and C. Karadi. Cat-a-cone: An interactive interface for specifying searches and viewing retrieval results using a large category hierarchy. In *Proceedings of SIGIR'97*, pages 246–254, Pilhadelphia, PA, USA, 1997.

[6] D. Kelleher. Spam filtering using contextual network graphs. Master's thesis, Computer Science Department, Trinity College, Dublin, Ireland, 2004.

[7] S. M. Kirsch. Social information retrieval. Diploma thesis in computer science, Rheinische Friedrich-Wihelms-Universität Bonn, Institut für Informatik III, Bonn, Germany, 22nd of November 2005.

[8] A. Kontostathis and W. M. Pottenger. Detecting patterns in the lsi term/term matrix. In *IEEE ICDM02 Workshop Proceedings, The Foundation of Data Mining and Knowledge Discovery (FDM02)*, 2002.

[9] A. Kontosthathis, W. M. Pottenger, and B. D. Davison. Assessing the impact of sparsification on lsi performance. In *Proceedings of the 2004 Grace Hopper Celebration of Women in Computing Conference*, Chicago, IL, USA, Oct 6-9 2004.

[10] R. R. Korfhage and K. A. Olsen. Information display: Control of visual representations. In *IEEE Workshop on Visual Languages*, pages 56–61, Kobe, Japan, 1991.

[11] S. Koshman. *Uasbility testing of a prototype visualization-based information retrieval system*. PhD thesis, University of Pittsburgh, 1996.

[12] R. C. T. Lee, J. R. Seagle, and H. Blum. A triangulation method for the sequential mapping of points from *N*-space to two-space. *IEEE Transactions on Computers*, 3(C-26):288–292, 1977.

Bibliography (continued)

[13] X. Lin. Graphical table of contents. In *Digital Libraries, DL'96*, pages 45–53, Bethesda, MD, USA, 1996.

[14] R. Mihalcea and P. Tarau. Textrank: Bringing order into texts. In L. Dekang and W. Dekai, editors, *Proceedings of EMNLP 2004*, pages 404–411, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[15] E. Morse and M. Lewis. Testing visual information retrieval methodologies case study: Comparative analysis of textual, icon, graphical, and "spring" displays. *Journal of the American Society for Information Science and Technology*, 53(1):28–40, 2002.

[16] S. E. Preece. *A Spreading Activation Network Model for Information Retrieval*. PhD thesis, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA, October 1981.

[17] H. Small. Visualizing science by citation mapping. *Journal of the American Society for Information Science*, 50(9):799–813, 1999.

[18] P. Snealth and R. R. Sokal. *Numerical Taxonomy*. W. H. Freeman, San Francisco, USA, 1973.

[19] J. A. Wise. The ecological approach to text visualization. *Journal of the American Society for Information Science*, 50(13):1224–1233, 1999.

[20] I. Lokuge and S. Ishizaki. Geospace: An interactive visualization system for exploring complex information spaces. In I. R. Katz, R. L. Mack, L. Marks, M. B. Rosson, and J. Nielsen, editors, *CHI95: Human Factors in Computing Systems, CHI 95 Conference Proceedings*, pages 409–414, Denver, Colorado, USA, May 7-11 1995. ACM, Addison-Wesley.

[21] G. Lane. Urban Tapestries: Wireless networking, public authoring and social knowledge. *Personal Ubiquitous Computing*, (7): 169–175, April 2003.

[22] E. L. Morse and M. Lewis. Why information retrieval visualizations sometimes fail. In *Proceedings of the 1997 IEEE International Conference on Systems, Man, and Cybernetics*, pages 1680–1685, Orlando, FL, USA, October 12-15 1997.