ARESOS Project

Reconstruction, Analyse et Accès aux Données dans les Grands Réseaux Socio-Sémantiques

Mission pour l'Interdisciplinarité du CNRS - Défi Masses de Données Scientifiques - MASTODONS

Patrick GALLINARI - UPMC Paris 6 - UMR 7606

Participants

- CAMS UMR 9557 - INSMI, EHESS, Paris
- CSI - UMR 7185 - INSHS, Ecole des Mines, Paris
- GIS Institut des Systèmes Complexes de Paris Ile-de-France, (Fédération de 16 instituts et universités), Paris
- IRISA, UMR 6074 - INS2I, IRISA, U. de Rennes 1
- IRIT, UMR 5505 - INS2I, U. Toulouse 3
- LATTICE, UMR 8094 - INSHS, ENS/ U. Paris 3
- LIG, UMR 5217 - INS2I, U. Joseph Fourrier, Grenoble
- LIP6, UMR 7606 - INS2I, U. Pierre et Marie Curie, Paris

Context

- Analysis of large socio-semantic networks
  - Production and diffusion of content on media
    - Human at the centre of the process
  - Caracterisation
    - Interactions
      - Individual + Social links
      - Structure of social interactions
      - multi-scale : micro, meso, macro, temporal
    - Dynamic of conversations and concepts
      - multi-scale
      - multi-sources

Diversity of information sources

TIME SCALES

Seconds	Google	> 180M Queries per hour
Minutes		> 400M tweet/d ; >540M SMS/d
Hours	BLOG	>100M blogs, +120.000 new blogs per day
Days	NEWS	~3200 AFP press release per day
Months	BETTER HEADGEAR THROUGH CHEMISTRY	~1,6M academic papers/year in the WoS (79.000 for FR, 216/d)

Project themes

- Analysis of content networks
  - Mainly textual
- 2 aspects
  - Representation and access to social content
    - Acquisition, indexation, querying, analysis, information flows, topic evolution, conversation following
  - Dynamicity : social-semantic structures and diffusion phenomena
    - Discovery of latent structures, morphogenesis, content diffusion
    - Co-evolution structure and semantic

Controversy

- Objective: who speak, about what, how?
- Identification of roles
  - Annotation platform
  - Linguistic analysis
- Topic identification
- Sentiment analysis
- Construction of socio-semantic networks
- Link analysis between documents, co-references
  - Latent models – role – themes
  - Dynamic evolution of thematic clusters clusters, individuals
- Social analysis
  - Social dynamics in corpora

Social IR

- Change of paradigm in IR
  - Identification and representation of social informations and social needs
    - Nature of information and needs
  - Relevence of information
- 3 axes
  - Information retrieval in microblogs
    - Detection of entities and relations
      - Individual, blogger communities, themes, opinions
    - Relevance of information - ranking
      - Authority, trust, temporality
  - Crowd Indexing
    - Indexation by crowd, social tagging,
  - Collaborative recommendation
    - Identify and score recommendations performed by user groups
    - Analysis of user groups
      - Temporal evolution of topics of interests

Dynamicity

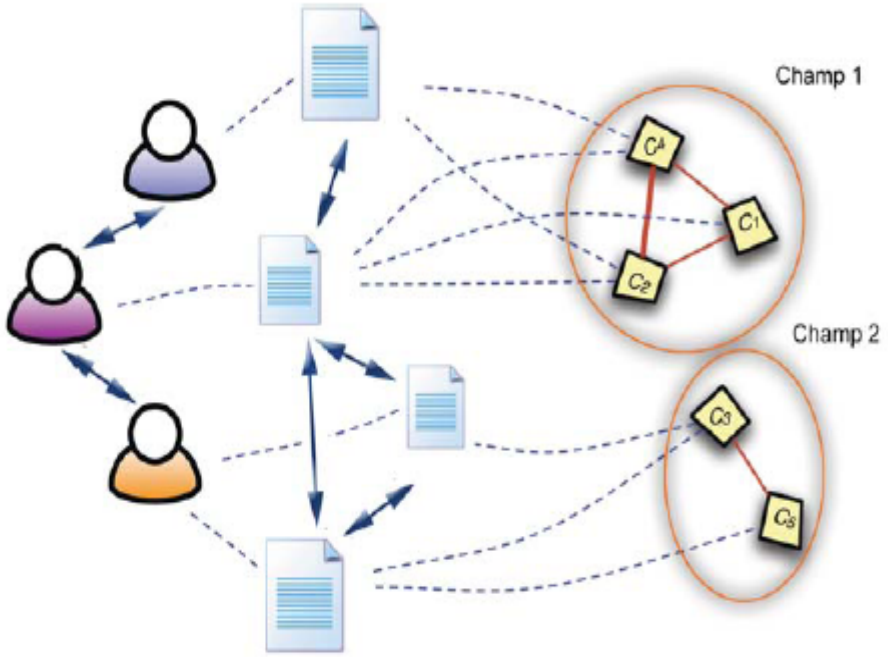
- Extraction of implicit structure
  - Joint analysis of semantic, thematic and social relations
- Modeling and morphogenesis
  - Modeling structure emergence
    - Thematic and social communities
  - Interactions between actors of the social dynamics
    - Temporal evolution of socio-semantic structure
    - Co-evolution individual and group dynamics
- Diffusion of content
  - Analysis and modeling of content diffusion
  - Characterization of roles
    - Influencers, initiators, gurus, ect

Phylomemetic reconstruction
Emergence of structures

D. Chavalarias, J.P. Cointet ISC-PIF

Multipartite and heterogeneous structures in digital media

Goal: modeling dynamic content production



Examples

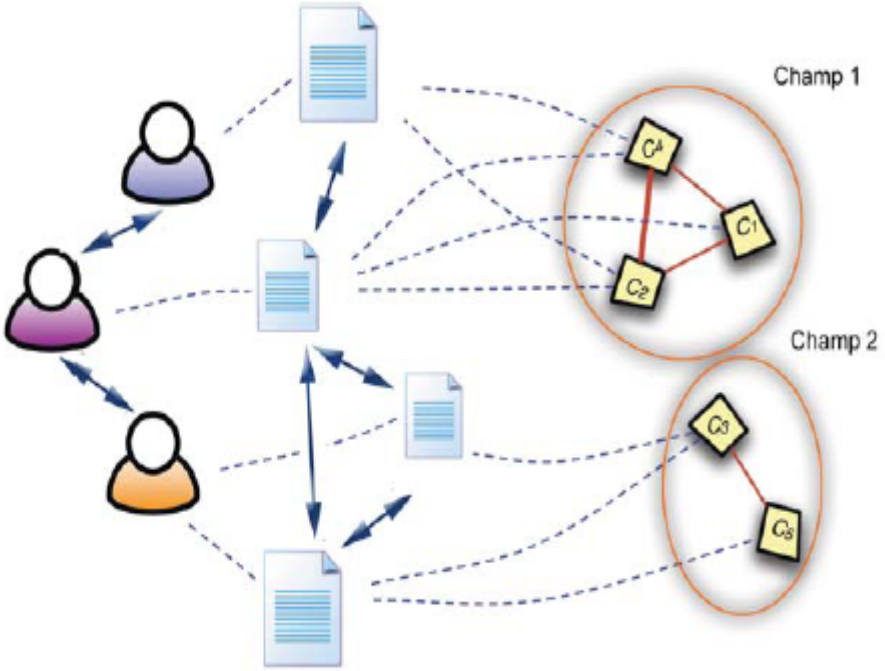
- Scientific papers, blogs, news, tweets & their keywords,
- Scientific papers, blogs, news, tweets & their authors ,
- Authors& their semantic profile.

Analysis at the micro (unit) /meso (community) / macro (multi-scale) structure

Each dimension can be viewed as dynamical substrates which could be used as background to study the other dimensions.

Multipartite and heterogeneous structures in digital media

Goal: modeling dynamic content production



Examples

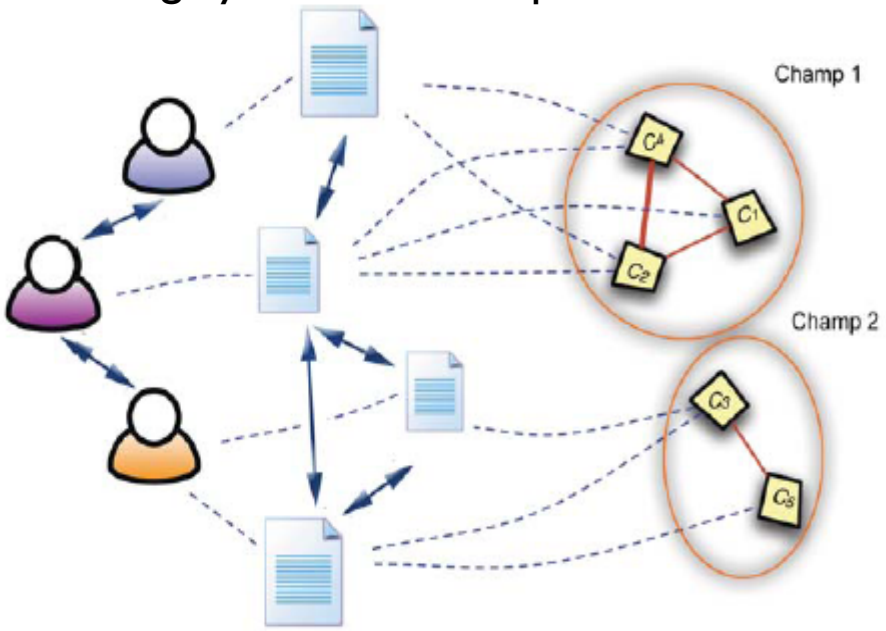
- Scientific papers, blogs, news, tweets & their keywords,
- Scientific papers, blogs, news, tweets & their authors ,
- Authors& their semantic profile.

Challenge

Perform the reconstruction of these dynamical multi-scale heterogeneous structures across different data sources with comparable methodologies.

Multipartite and heterogeneous structures in digital media

Goal: modeling dynamic content production



Examples

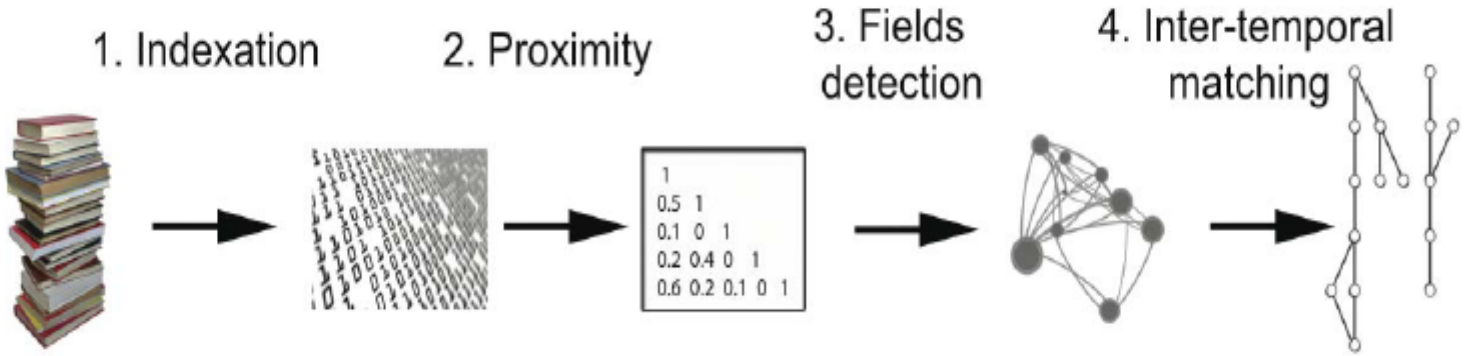
- Scientific papers, blogs, news, tweets & their keywords,
- Scientific papers, blogs, news, tweets & their authors ,
- Authors& their semantic profile.

Questions

- Which are the morphogenesis rules in each media ? What are the characteristic dynamical patterns ? What can be modelled and predicted ?
- How the knowledge issued from the phenomenological reconstruction of socio-cultural dynamics is articulated with questions stemming from Social Sciences and Humanities ?

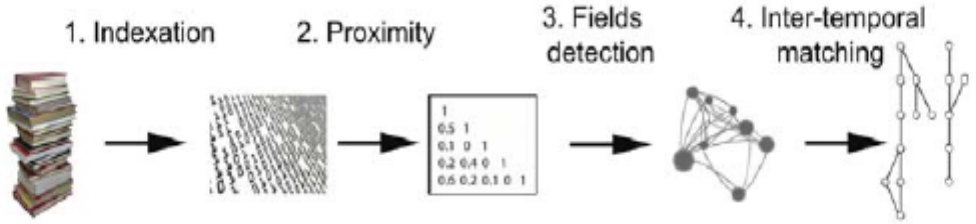
Example of reconstruction of the semantic landscape dynamics

Goal: modeling dynamic content production

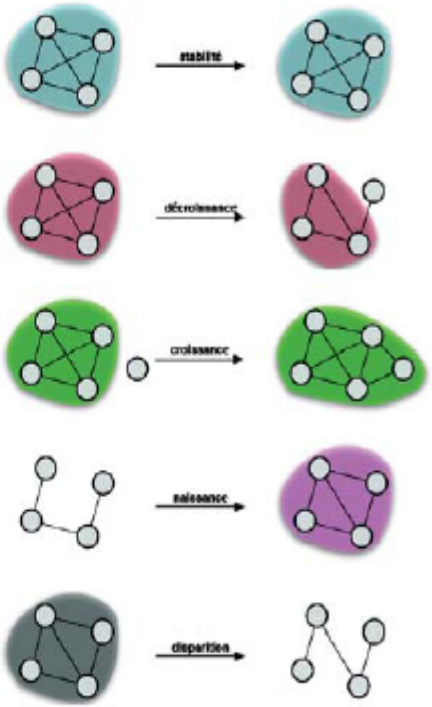


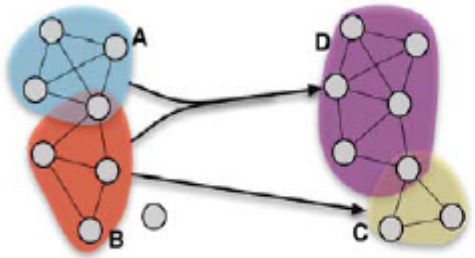
1. **Semantic analysis and indexation** on large corpora,
2. **Computation of proximities between terms**
3. **Topic detection :** community detection algorithms on directed, weighted graphs ; with overlap Time dependent clustering
4. **Phylogeny reconstruction** of topics

Topic dynamics ("Meso" level)

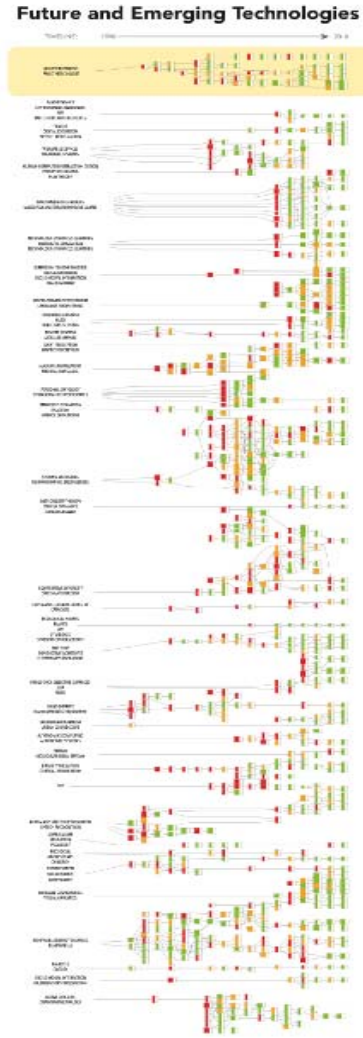


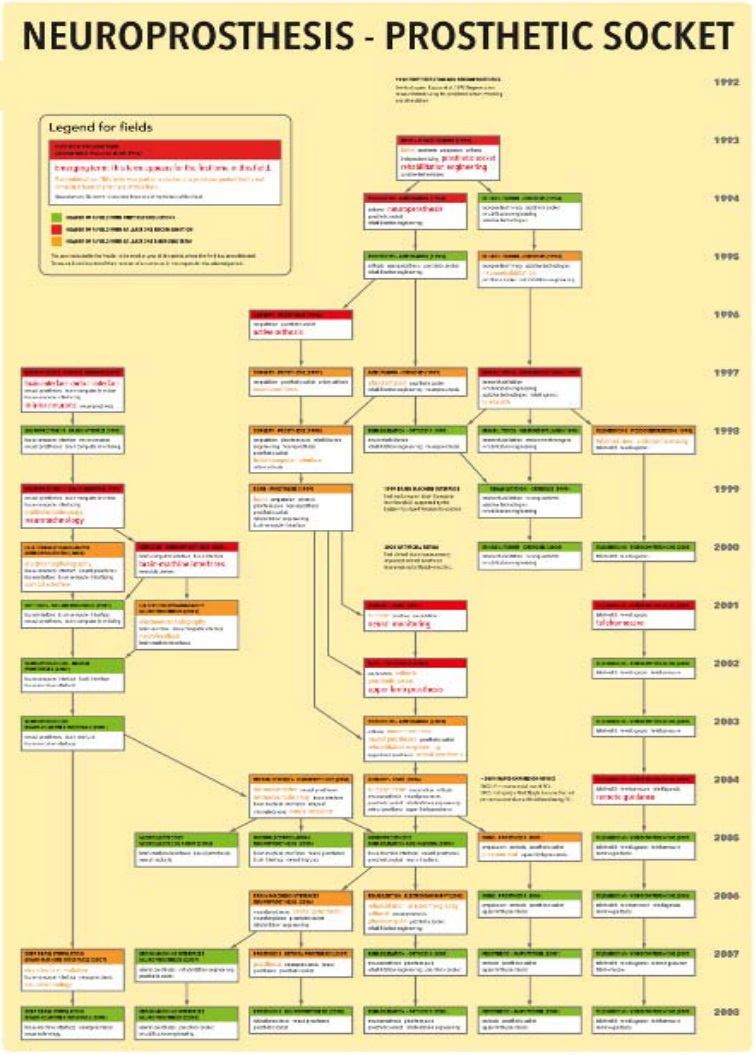
Clusters at t Clusters at t + T



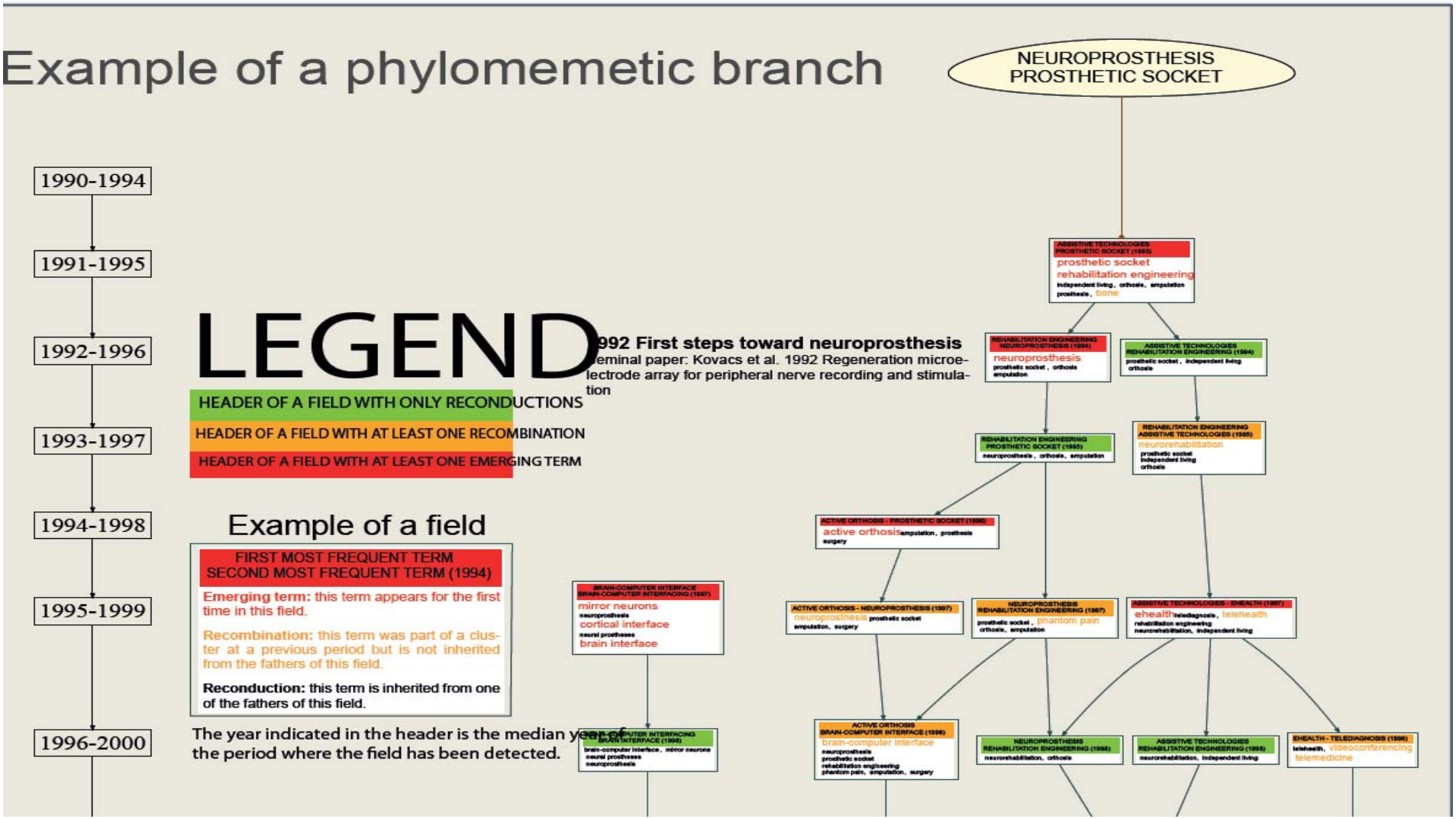


FET Open Phylomemy



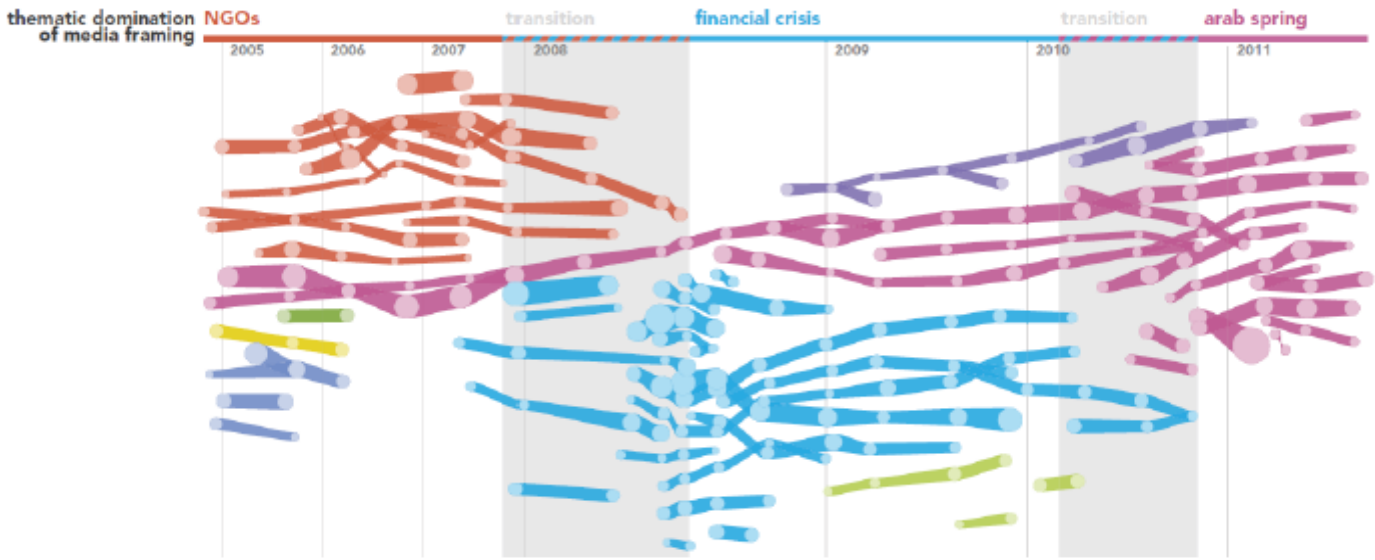


Detail: Neuroprosthesis



Topic dynamics - Media sphere

Monitoring food crisis worldwide



2004 - 2008
food insecurity in humanitarian crises
caused by natural disasters(floods/drougts), or wars and conflicts
food security & agricultural policies

2008 - 2010
food insecurity & poverty: the '08 financial crisis
food riots spreading, increased number of undernourished people, management of food as a global issue by international institutions

2010 - 2011
food insecurity: a cause of social unrest
from global crisis to local consequences
food insecurity: an health problem
continuity of infant malnutrition issues

Chavalarias, Cointet, Cornilleau, Duong, Mogoutov, Villard, Roth, Savy 2011. http ://pulseweb.veilledynamique.com

Social Information Retrieval

(M. Boughanem et al. IRIT, Toulouse)

Social Media

- Blogging
  - Blog (Blogger, Technocrati)
  - Twitter (micro-blogging)
- Outil de publication
  - Wiki
- Social Networking
  - Facebook, MySpace, Classmates
  - LinkedIn, Plaxo, Xing
- Bookmarking sites
  - Del.icio.us, blogmarks, dogear
- Folksonomy (Social tagging, …)
  - Flickr, Photobucket, YouTube
- Forum de discussion
  - PhpBB, Skype




Rich, Big and Fast

- Rich and diverse
  - Textual, Multimedia (image, videos, etc.)
  - Billions of connections
  - Behaviours
  - Preferences
  - Opinions
  - Comments
  - Trends...

Social search

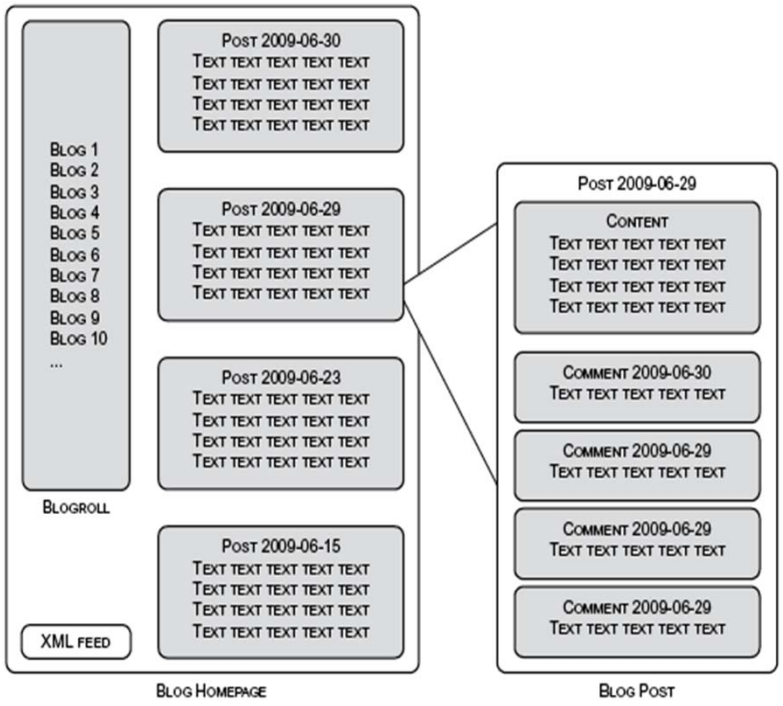
- « *Social search : how social interactions and social data can enhance existing information-seeking experiences, as well as enable new information retrieval scenarios.*

- *The different models of social search, including:*
  - *1) social data as new information to be searched*
  - *2) use of social data to augment search*
  - *3) social interaction and collaboration as part of the search process*

Blog search

- Find blog to subscribe to for topic X
  - Site should contain many relevant posts and be topic focused

Overview of the TREC 2006 blog track (2006),
(Seo &Croft CIKM 2008),
(Mishne & de Rijke ECIR 2006),
(M.A. Hearst et al. SSM 2008)



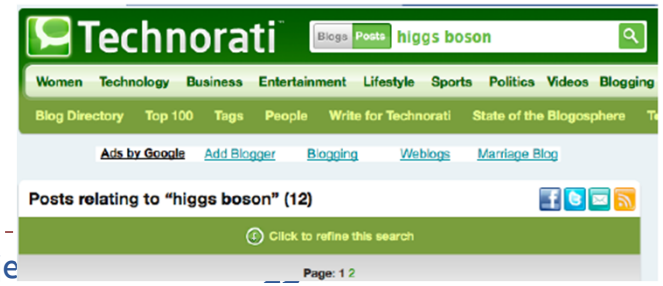
Foundations and Trends® in
Information Retrieval
Vol. 6, No. 1 (2012) 1–125

DOI: 10.1561/1500000026

now
the essence of knowledge

Information Retrieval on the Blogosphere

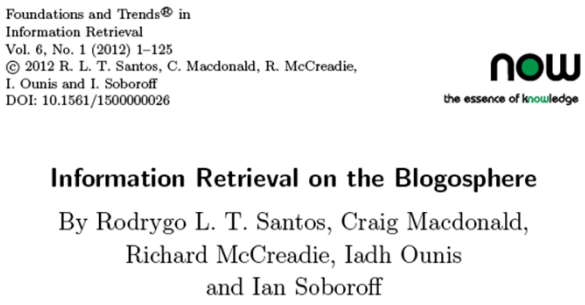
By Rodrygo L. T. Santos, Craig Macdonald,
Richard McCreadie, Iadh Ounis
and Ian Soboroff



Opinion retrieval (sentiment analysis)

- Find relevant and opinionated (positive or negative) document (post, tweet, sentence, …) about a given topic
  - Machine learning, lexicon-based (Wu, 2008] [Thewall, 2009] [Mishne, 2006] [Agrawal, 2003, M. Missen et al]

- Opinion summarization (aggregation)
  - Product review mining;
  - Tracking sentiments toward topics over time;
  - Prediction (election outcomes, market trends)?



Expert search

- Find experts on a given topic (*for being asked for questions, assigned some role or job in an organizational setting*).

Foundations and Trends® in
Information Retrieval
Vol. 6, Nos. 2–3 (2012) 127–256

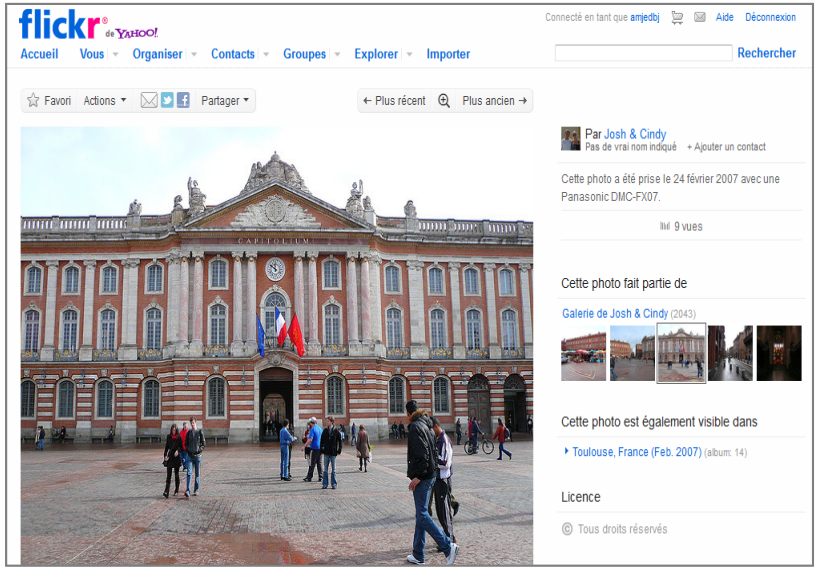
DOI: 10.1561/1500000024

now
the essence of knowledge

Expertise Retrieval

By Krisztian Balog, Yi Fang, Maarten de Rijke,
Pavel Serdyukov and Luo Si

Social tagging



- Many users annotate photos)
  - Capitole, Toulouse
  - Mairie de Toulouse, place du capitole
  - Toulouse

- Searching for items using tags
- Tag cluster visualization

Sigurbjornsson and van Zwol, WWW 2008

Social recommendation

- Presenting most relevant content “suggested/advised/liked ” by other users (your friends)
  - Item-based method (deployed by Amazon)
    - Similaritry and trust *[Jamali, 2009] [Ma, 2009] [McDonald, 2009]*
  - User-based method
    - Collaborative filtering *[Konstas, 2009] [Siersdorfer, 2009]*

Focus: Microblog search

> “ Microblogging is a new form of **communication** [….] that enables users to **broadcast** and **share information** about their **activities**, **opinions** and **status**. [Java & *al.2007]*.

- Microblog post
  - Short (140 characters)
  - Real-time
  - Social motivation

twitter

+400 millions	Publications /day
+500 millions	User accounts
2 Billion	Search queries/day

Tweet, URL, hashtag, Reply, Mention , RT …

Retweet

Barack Obama @BarackObama — 7 Nov 2012

Four more years. pic.twitter.com/bAJE6Vom

Mention

Retweeted by **Jack Dorsey** — 809,104 RETWEETS 301,873 FAVORITES

David Cameron @David_Cameron — 7 Nov 2012

Warm congratulations to my friend @BarackObama. Look forward to continuing to work together.

Reply

2,239 RETWEETS 480 FAVORITES

Alicia Keys @alciciakeys — *7 Nov 2012*

@BarackObama WE did it!!!

Hasht

View conversation — **467** RETWEETS **242** FAVORITES

Twitter Government @gov — *7 Nov 2012*

With 20 million tweets, Election Day just became the most event in US political history. #election2012

URL (photo, video, blog, etc)

Barack Obama @BarackObama

Four more years. pic.twitter.com/bAJE6Vom

View photo — **809,104** RETWEETS **301,873** FAVORITES

Microblog IR

- Microblog IR tasks
  - Person search (to follow)
  - Trend extraction
  - Event detection and tracking
  - Opinion search
  - Microblog (e.g. tweet) search

Microbog search

- Finding the most relevant tweets for a given topic

- Search motivations
  - access to concise and credible information
  - access to fresh and real-time news (traffic jam, down services…)
  - follow an event
  - collect opinions and public sentiments

Search on Twitter (Teevan et al. WSDM 2011)

	Web Search	Twitter Search
Query length (chars)	18.80	**12.00**
Query length (words)	3.08	**1.64**
Is a celebrity name	3.11%	**15.22%**

Challenges

- Real-time indexing and searching
  - Efficient indexing in order to provide fast results
    - Index stream data
      - Distributed indexing (Bush et al ICDE'12)
      - Selective indexing (Chen et al SIGMOD'11
        - Limit the number of tweets to be indexed (only 20% of queries represent 80% of user requests
  - Effective ranking in order to return relevant results
    - Relevance model (factors to be used to handle relevance)
      - Content features: Tweets, hashtags, URL, @adr, RT,
      - Timestamp: Freshness of the tweet
      - Social Features: followership, Retweets, Reply, Mention, sentiment

Salient features

Topical features

Exact term matching

Tweet popularity

Hashtags popularity

Topic as hashtags

Syntactical features

Hashtags presence

URL presence

Is-reply

Tweet length

Social features

Number of tweets

Mention

Retweet frequency ?

Number of followers of an author?

Temporal feature :

Query time vs. tweet time

Semantic features :

Expand the query

Hashtag expansion

"#TextAndDrive" become "Text and Drive"

Ranking model (Ben jabeur et al WI 2012)

- Bayesian network model for combining
  - Topical relevance RSV(Q,t)
  - Timestamp f(Qtime, ttime)
  - Social relevance: microblogger (influence + expertise)

$$Rel(Q,t,G) = \alpha RSV(Q,t) * f(Q^{time}, t^{time}) + (1-\alpha) S(Q,u_t,G)$$

Q: query
t : Tweet
G : Social network
u_t : microbloger

$$RSV(Q,t) = P_{dir}(Q,t) = \prod_{w \in Q} \frac{P_{dir}(w|t) + \mu * P(w|C)}{\mu + |t|}$$

$$f(Q^{time}, t^{time}) = \ker nel(Q^{time}, t^{time})$$

- Tests on TREC Micro-Blog 2011-2012

Conclusion

- Social search → very active area
- → several challenges
  - Searching ephemeral data
  - Mining social data to enhance a search
  - Aggregating social data

- Merci
- http://mastodons.lip6.fr/