ERIC
ENTREPÔTS, REPRÉSENTATION & INGÉNIERIE des CONNAISSANCES

UNIVERSITÉ LUMIÈRE LYON 2
UNIVERSITÉ DE LYON

Extracting and evaluating topics. *CommentWatcher*, an online forum analysis tool

Marian-Andrei RIZOIU

ERIC Laboratory, University Lumière, Lyon, France

WSC 2013 *August 26th, 2013*

Hong Kong, China

Part I.

Topic extraction, topic labeling and semantic-aware topic evaluation

Dataset: Collection of natural language texts, usually issued from the internet

Difficulties:

- ➜ great volumes of data
- ➜ need to summarize the main "ideas": *the topics*
- ➜ most of the literature evaluates extracted topics using statistical measures, completely short-circuiting the semantics

ex. the perplexity index [WAL09]

Learning tasks:

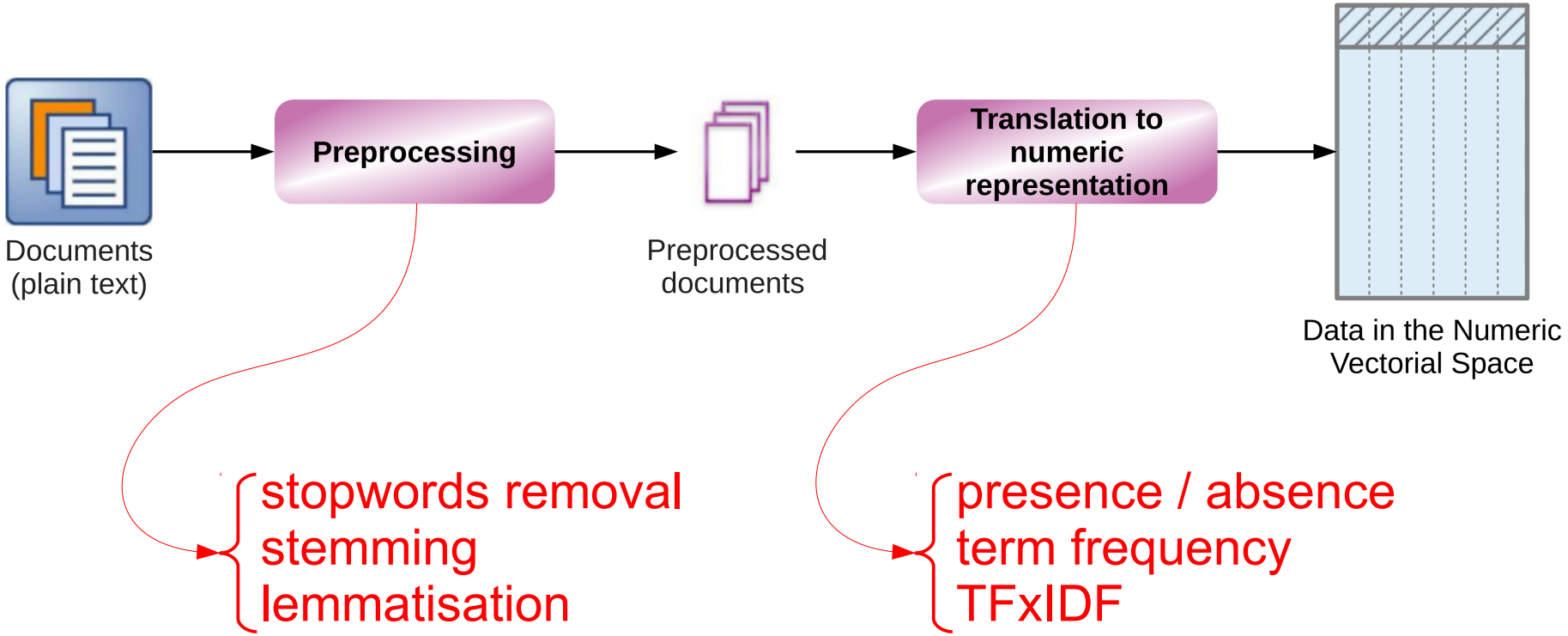
- topic extraction
- labeling topics with names which are comprehensible for a human being
- using semantic knowledge in the evaluation of the topics

Applied dimension:

- real, strong demand from scientists in **Social Sciences and Humanities** *(Sociology, Psychology, Linguistics, History, etc.)*
- implemented in the forum analysis web-based platform **CommentWatcher**

Proposed solution (1): An alternative to graphical models (*e.g.*, LDA [BLE03]): textual clustering

Prerequisites: Translating the data into a vectorial numeric space: “bag-of-words”



Proposed solution (2): Topic extraction and labeling

I. Extract topics using an overlapping textual clustering algorithm

OKM [CLE08] An extension of Kmeans which authorises documents to belong to multiple clusters

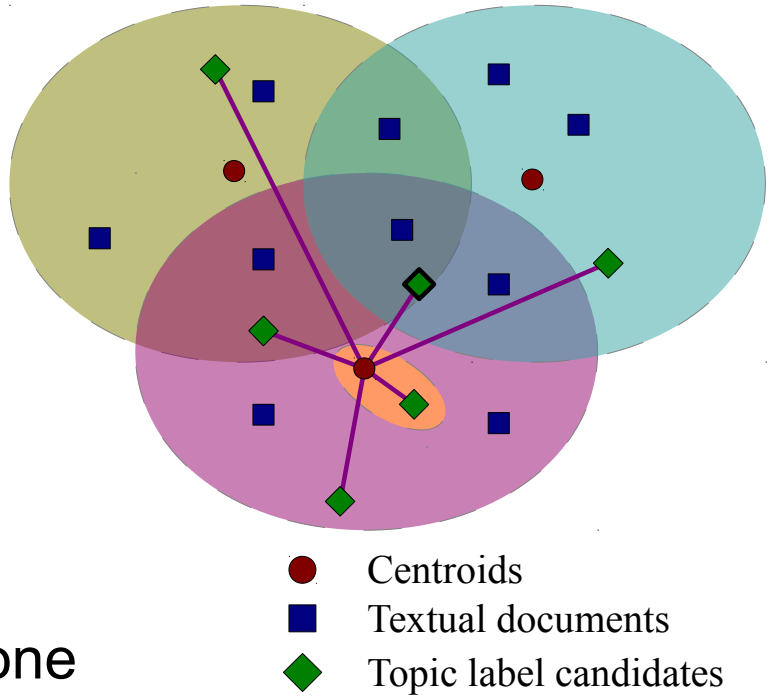
- ➜ construct the overlapping partition
- ➜ Centroids are abstractions of their cluster: *topics*

II. Label topics using frequent expressions

- ➜ extract frequent complete expressions from the original text

  suffix array [MAN93].

- ➜ Inject the expressions in the document space as pseudo-documents
- ➜ Calculate similarity and chose the closest one



Proposed solution (3): Evaluate the semantic cohesion of topics

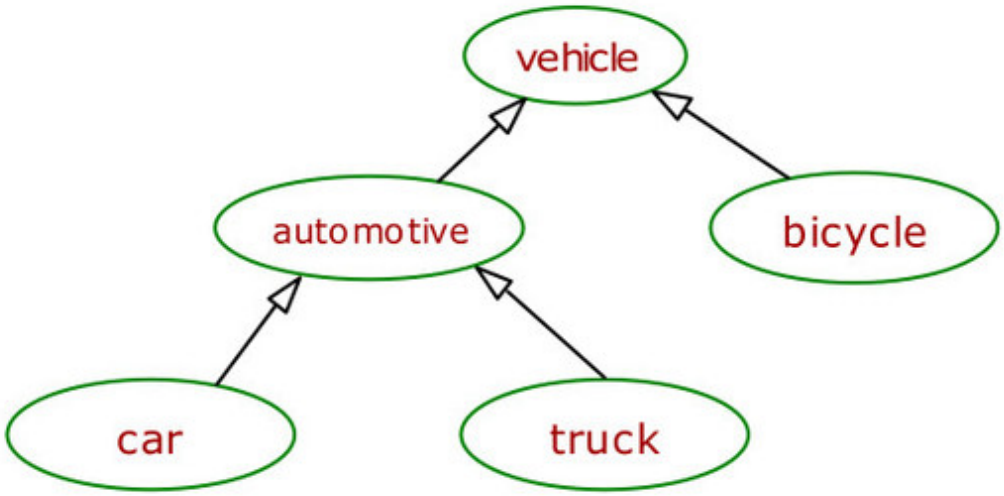
Underlying assumption: Statistical measures do not completely emulate human judgement of topics [CHA09].

The idea: Map a statistic frequency distributions (a topic) to a semantic structure

Use the most pertinent terms attached to a topic

WordNet [MIL95]

- concept hierarchy
- a concept groups together sets of synonyms
- Polysemy: a word has multiple meanings and a concept is a sense of a word



Proposed solution (3): Evaluate the semantic cohesion of topics

Topic alignment: Determine the most specific subtree which contains at least a sens for each of the most representative words of the topic

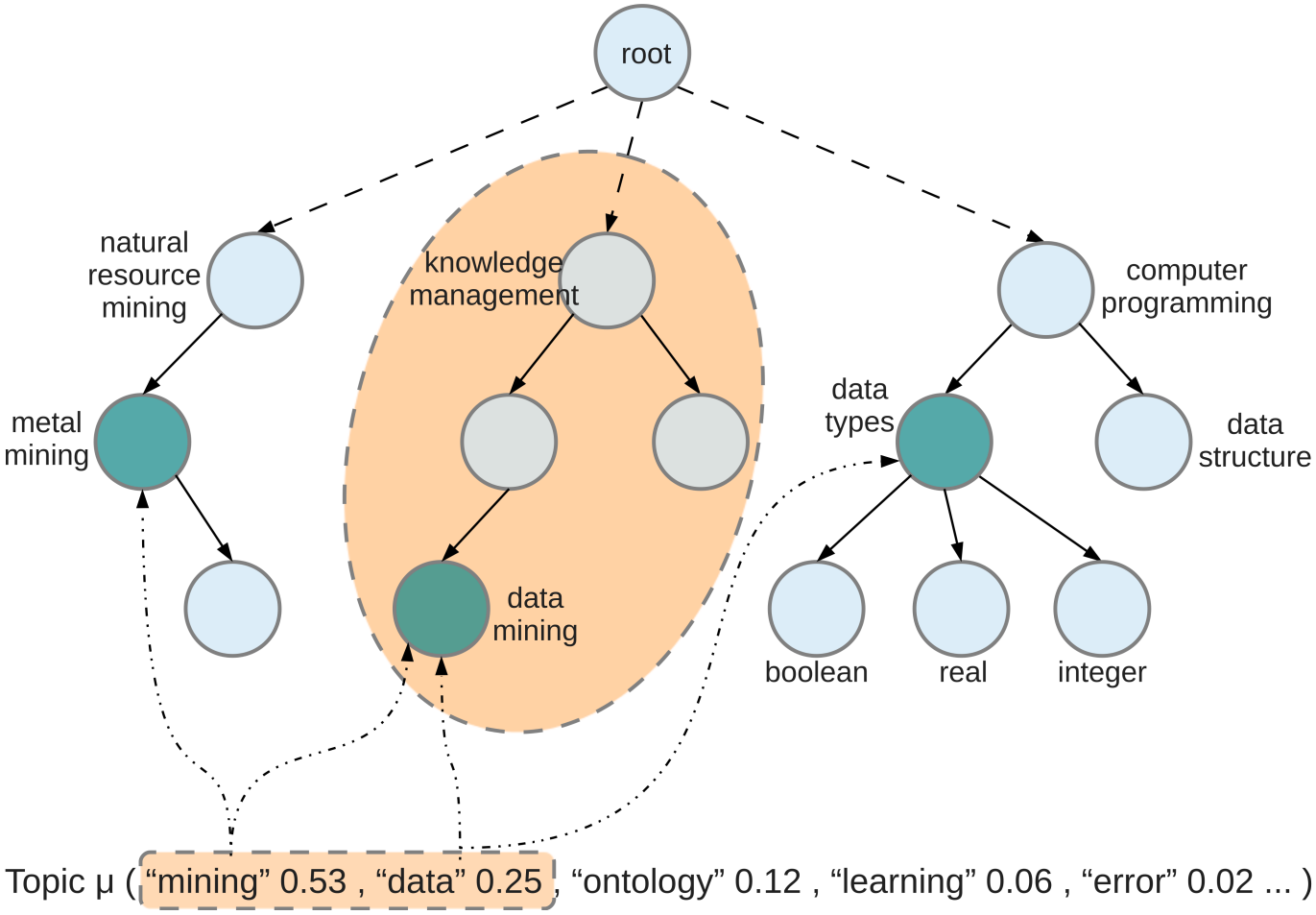
Measures:

coverage

specificity

Topic fitness function:

$$\varphi(\mu, c) = \omega_{spec}\, spec(\mu, c) + \omega_{cov}\, cov(\mu, c)$$



Experiments and results:

Reuters, **Suall11**

French web forum “Y a-t-il trop de commémorations en France?”, sur www.liberation.fr

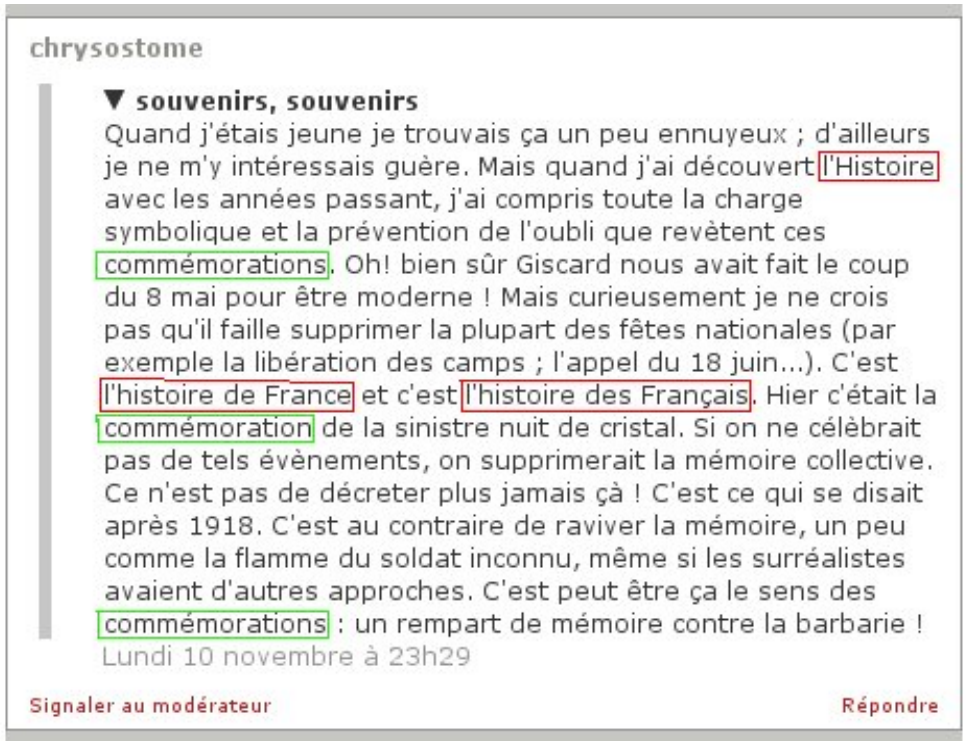
Economic dataset extracted from the site of Associated Press

```
--> Iteration no 11:
---> Objective function value: 189.154
---> Partitions:
----> Cluster 0 [101]: .....
----> Cluster 1 [90]: .....
----> Cluster 2 [128]: ..... texte_81 .....
----> Cluster 3 [192]: ..... texte_81 ....

Result - Cluster description:

-> Centroid[0]: "jours fériés"
-> Centroid[1]: "travailler plus pour gagner"
-> Centroid[2]: "commémoration"
-> Centroid[3]: "histoire de france"
```

Example of output of the clustering-based topic extraction software included in **CommentWatcher**



chrysostome

▼ **souvenirs, souvenirs**

Quand j'étais jeune je trouvais ça un peu ennuyeux ; d'ailleurs je ne m'y intéressais guère. Mais quand j'ai découvert l'Histoire avec les années passant, j'ai compris toute la charge symbolique et la prévention de l'oubli que revètent ces commémorations. Oh! bien sûr Giscard nous avait fait le coup du 8 mai pour être moderne ! Mais curieusement je ne crois pas qu'il faille supprimer la plupart des fêtes nationales (par exemple la libération des camps ; l'appel du 18 juin...). C'est l'histoire de France et c'est l'histoire des Français. Hier c'était la commémoration de la sinistre nuit de cristal. Si on ne célèbrait pas de tels évènements, on supprimerait la mémoire collective. Ce n'est pas de décreter plus jamais çà ! C'est ce qui se disait après 1918. C'est au contraire de raviver la mémoire, un peu comme la flamme du soldat inconnu, même si les surréalistes avaient d'autres approches. C'est peut être ça le sens des commémorations : un rempart de mémoire contre la barbarie !

Lundi 10 novembre à 23h29

Signaler au modérateur Répondre

Document “texte_81” on the website of the forum

Experiments and results:

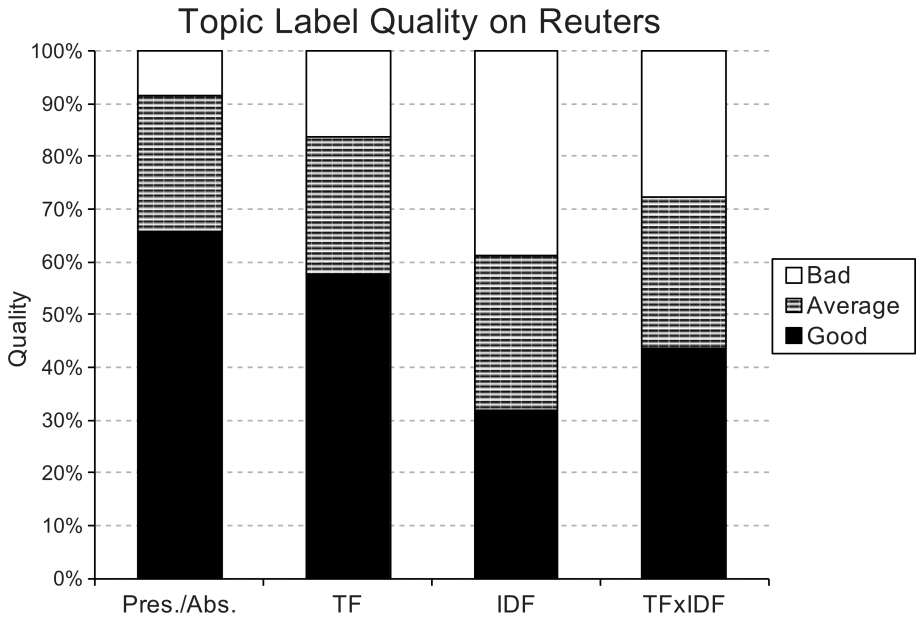
Reuters, **Suall11**

French web forum “Y a-t-il trop de commémorations en France?”, sur www.liberation.fr

Economic dataset extracted from the site of Associated Press

Experimental protocol :

Expert-based, inspired from the literature [CHA09]



Evaluation of topic labels

Dataset	$\overline{hit}_+$	$\overline{hit}_-$	Avantage rel. $\overline{hit}$
AP	**0,69**	0,65	6,93 %
Suall11	**0,75**	0,59	28,55 %

Evaluation of topic alignment to concept subtrees

Part II.

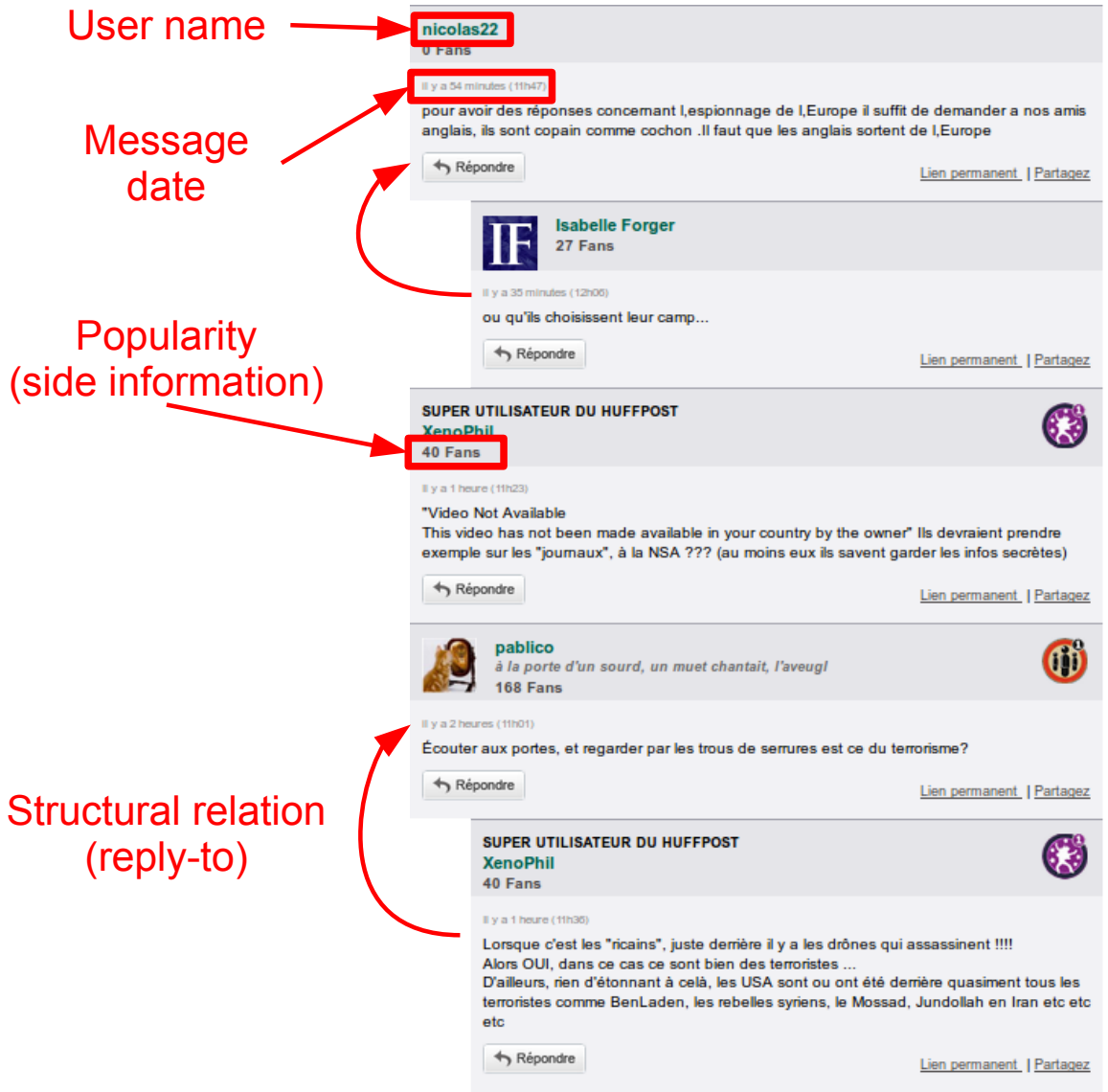
CommentWatcher – a web-based platform for analyzing online forum discussion

Applied work – CommentWatcher

Discussion forum analysis platform

Difficulties:

- Most existing tools do not treat the social network aspect of forum data [AME12, GUI13]
- Lack of benchmark forum datasets
- The structure of the hosting websites changes constantly
- Licensing problems with the content of forums



General objective:

two types of users

Forum analyst: study the discussion topics and user interaction

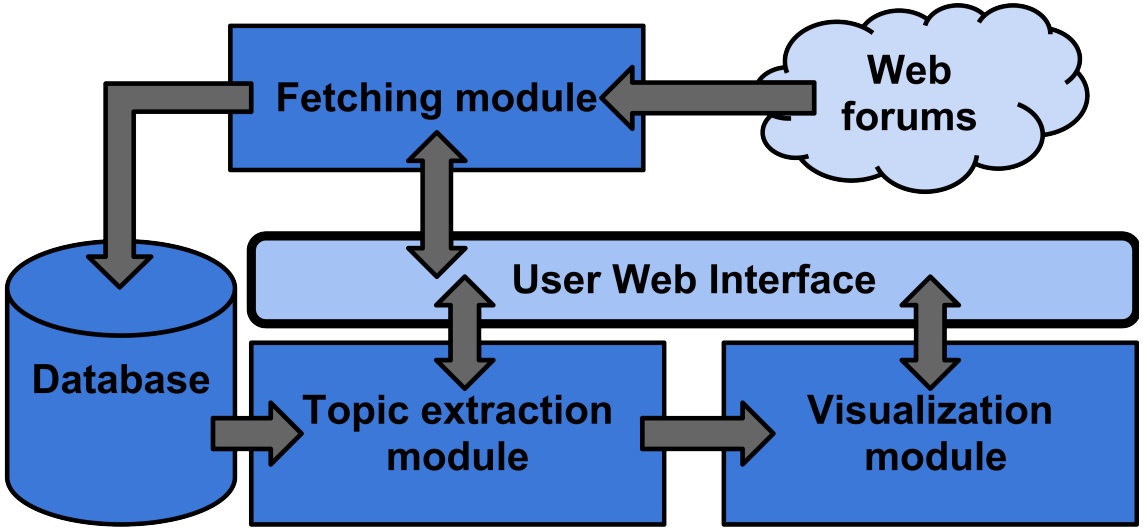
Researcher: construct discussion forum datasets, analyze the evolution of discussion topics

Our proposal: *CommentWatcher*

Plateforme Web opensource (GPLv3)

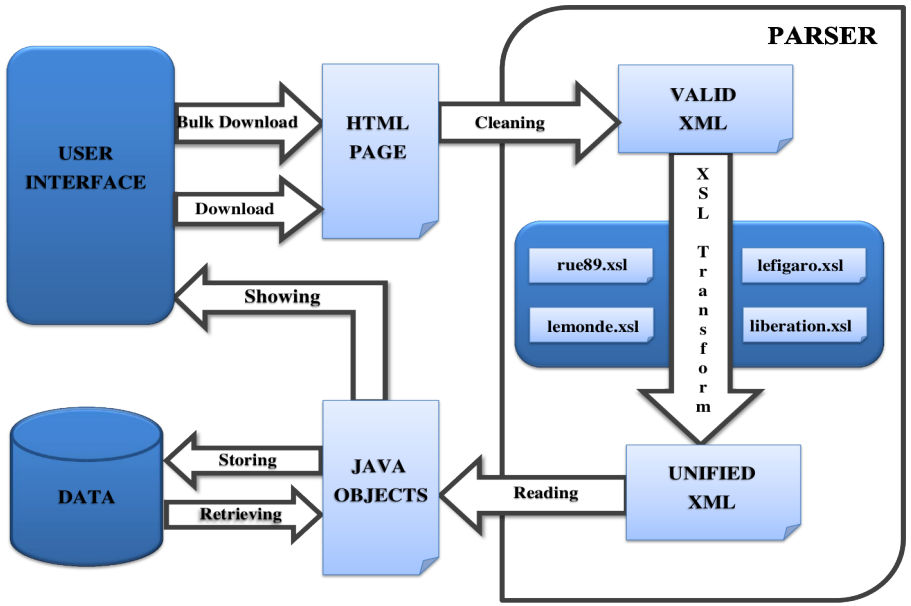
4 tasks:

- Retrieve data from internet (meta-parser)
- Topic extraction
- Topic visualization as an expression cloud and temporal evolution
- Visualization of the underlying social network



Module I. Data fetching

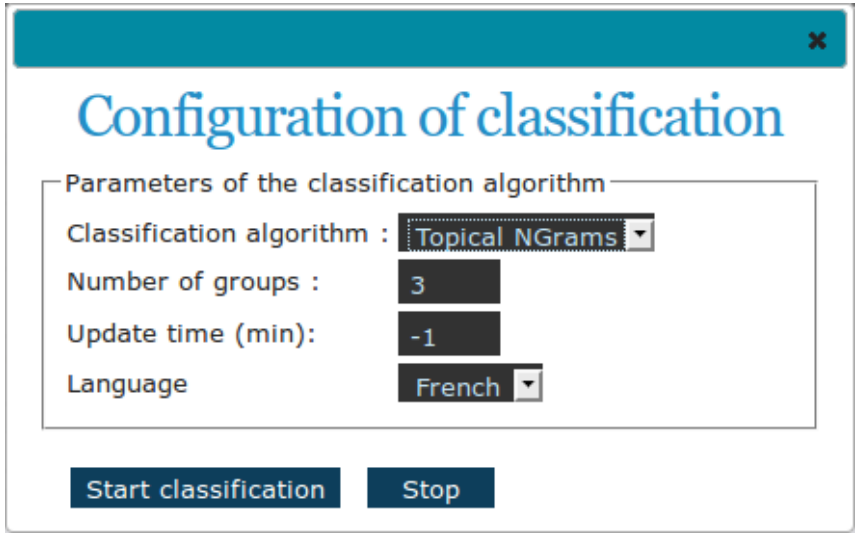
- Meta-parser, independent from the structure of web pages
- Support for new websites via definition files
- Search for supported forums via web querying and support for “mass fetching”



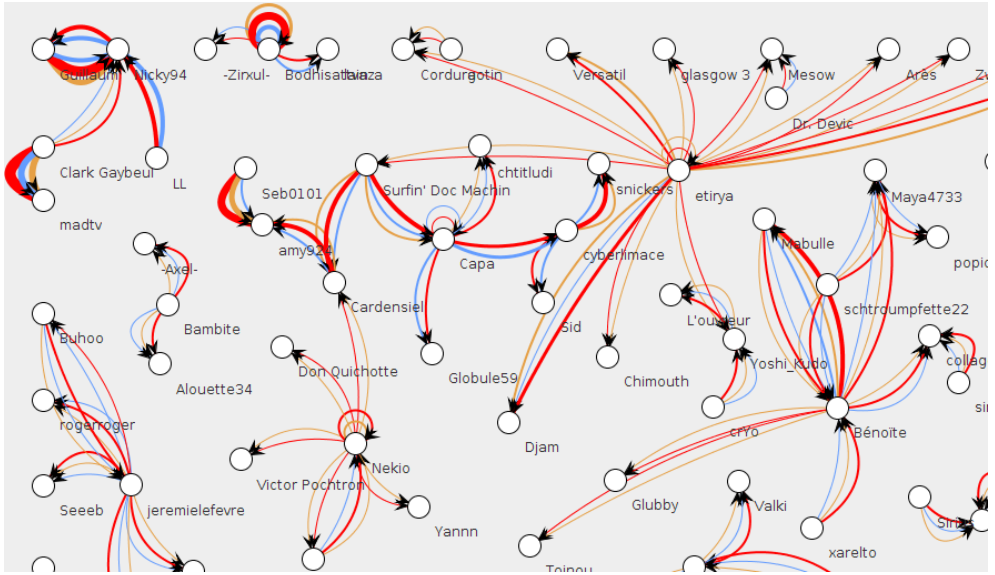
Module II. Topic extraction

3 algorithmes supportés :

- Topical Ngrams (Mallet [MCC02])
- CKP [RIZ10]
- Dynamic Topic Models [BLE06] **(in development)**



Module III. Visualizers



- Expression cloud for each topic
- Temporal evolution by forum and by website
- Evolution of the popularity of a topic

- Social network modelled as a multigraph
- **Vertexes:** the users; **Arcs:** the messages associated with the topics
- Based on the citation relation

Video Demonstration

Presentation website : http://mediamining.univ-lyon2.fr/commentwatcher

Conclusion

- A framework for topic extraction and labeling using a textual overlapping clustering
- Topic semantic evaluation using a topic – concept mapping
- *CommentWatcher* – an opensource web-based platform for discussion forum analysis

Perspectives

- Mapping hierarchies of topics (hLDA) on hierarchies of concepts
- Adding support for temporal topic extraction and adapted visualization
- Integrating the calculation of social network measures
- Evolving the visualization from the current client side (applet) towards server side

Bibliographie

[BLE03] David M. Blei, Andrew Y. Ng and Michael I. Jordan, Latent dirichlet allocation (2003), in: The Journal of Machine Learning Research, 3(993--1022)

[WAL09] Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov and David Mimno. Evaluation methods for topic models. In International Conference on Machine Learning, Proceedings of the 26th Annual, pages 1105–1112. ACM, 2009.

[MAN93] Udi Manber and Gene Myers. Suffix arrays: A new method for on-line string searches. SIAM Journal on Computing, vol. 22, no. 5, pages 935–948, 1993.

[CLE08] Guillaume Cleuziou. An extended version of the k-means method for overlapping clustering. In Pattern Recognition, 2008. ICPR 2008. 19th International Conference on, pages 1–4. IEEE, 2008.

[MIL95] George A. Miller. WordNet: a lexical database for English. Communications of the ACM, vol. 38, no. 11, pages 39–41, 1995.

[CHA09] Jonathan Chang, Jonathan Boyd-Graber, Sean Gerrish, Chong Wang and David M. Blei. Reading Tea Leaves: How Humans Interpret Topic Models. In Advances in Neural Information Processing Systems, Proceedings of the 23rd Annual Conference on, volume 31 of NIPS 2009, 2009.

[AME12] S. Amer-Yahia, S. Anjum, A. Ghenai, A. Siddique, S. Abbar, S. Madden, A. Marcus, and M. El-Haddad. Maqsa: a system for social analytics on news. In SIGMOD '12, pages 653–656, 2012.

[GUI13] A. Guille, C. Favre, H. Hacid, and D. Zighed. Sondy: An open source platform for social dynamics mining and analysis. In SIGMOD '13, 2013.

[MCC02] A. K. McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

[RIZ10] M.-A. Rizoiu, J. Velcin, and J.-H. Chauchat. Regrouper les données textuelles et nommer les groupes à l'aide des classes recouvrantes. In EGC '10, page 561, 2010.

[BLE06] David M Blei and John D Lafferty. Dynamic topic models. In Proceedings of the 23rd international conference on Machine learning, pages 113–120. ACM, 2006.

[MUS11] Claudiu Musat, Julien Velcin, Stefan Trausan-Matu and Marian-Andrei Rizoiu. Improving topic evaluation using conceptual knowledge. In International Joint Conference on Artificial Intelligence, Proceedings of the Twenty-Second, volume 3 of IJCAI 2011, pages 1866–1871. AAAI Press, 2011.