Extracting user interests from search query logs: a clustering approach

Lyes Limam, David Coquil, Lionel Brunie, Harald Kosch

Presented by: David Coquil

David.Coquil@uni-passau.de

DEXA-TIR Workshop, 30.08.2010

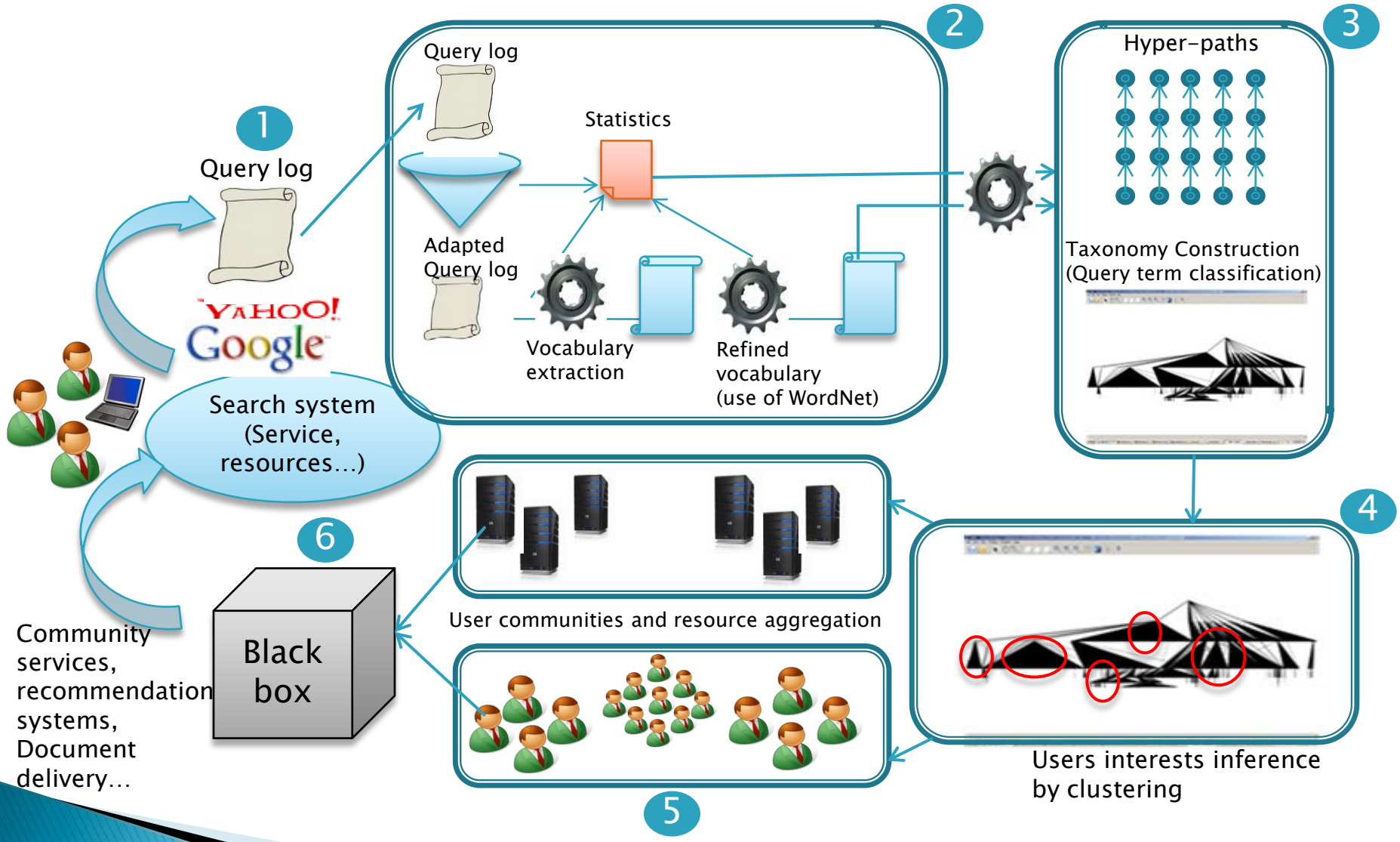
Introduction (1)

- User-centric systems
  - Design stage
  - Production stage
- Needs of online user-centrism
  - Gain knowledge from user interactions
- User logs analysis

Introduction (2)

- Query logs analysis
- Semantic analysis
- Textual search queries analysis
  - Semantically: identifying user interests
  - Technically: a query terms clustering problem

Framework for usage analysis



What do we need in our method ?

- Restructure the query logs to enable quantifying terms relationships
  - External source of semantic information
- Query terms clustering algorithm
- Semantic distance

WordNet as external source of semantics

- (English) WordNet
  - Large number of synsets
  - Hypernymy/(IS-A) relations

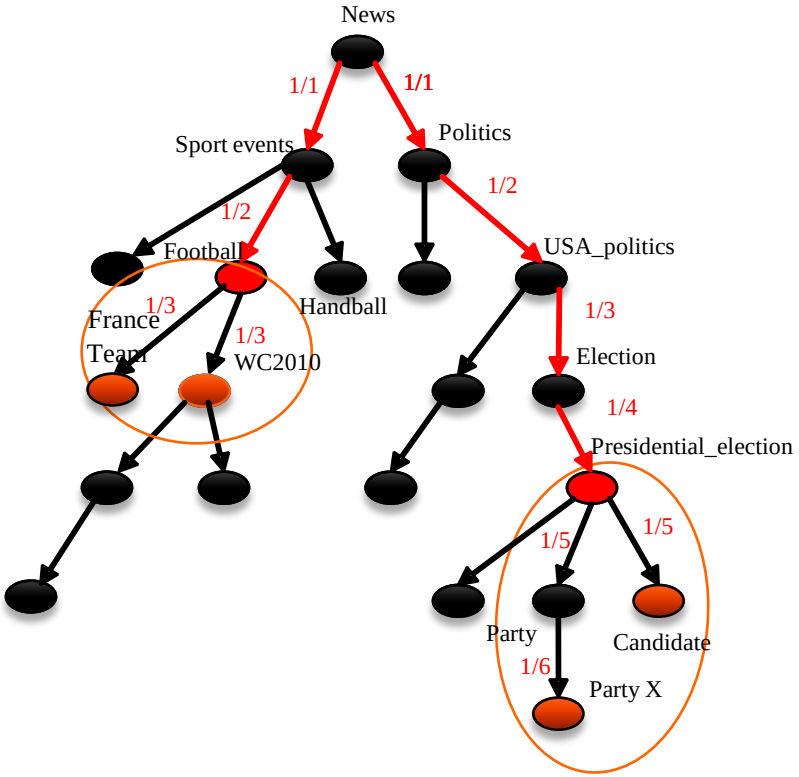
- Representation of the logs as a hierarchical structure

Preliminary phases

- Preprocessing
  - Elimination of unusable queries
  - Stop words
- Taxonomy construction process
  - Vocabulary
  - Hypernymy paths
  - Virtual nodes

Query term classification (Keywords Taxonomy)

- Global semantic representation of the log
- Defines a metric that measures the semantic distance between the terms
- A base for analysis
  - query terms clustering process



Semantic distance function

- The distance function is defined as follows:

- *G(V,E) a tree structure*
  - *V the set of terms*
  - *E the set of edges that models the relationships* ***term1"is-a" term2***

- *Let "L" be a function which returns the level of an element*
- *The weight function "W" is defined on "E" as :*
-

$$\forall (u,v) \in E \,/\, u"is-a"v : W(u,v) = 1/L(v)$$

- *Let P = {e_1, …, e_n} the set of edges in the path (unique) between x and y : (x,y)∈ V^2*
- *The distance function "D" is defined on V^2 as :*

$$\forall (x,y) \in V^2 : D(x,y) = \sum_{i=1}^{n} W(e_i)$$

Clustering Algorithm

- Groups terms whose all the distances are less than a threshold
- The clusters are constructed by pruning
  - The construction starts from the bottom
- The algorithm :
  - Is deterministic
  - Its complexity is O(*n*), where *n* the number of nodes

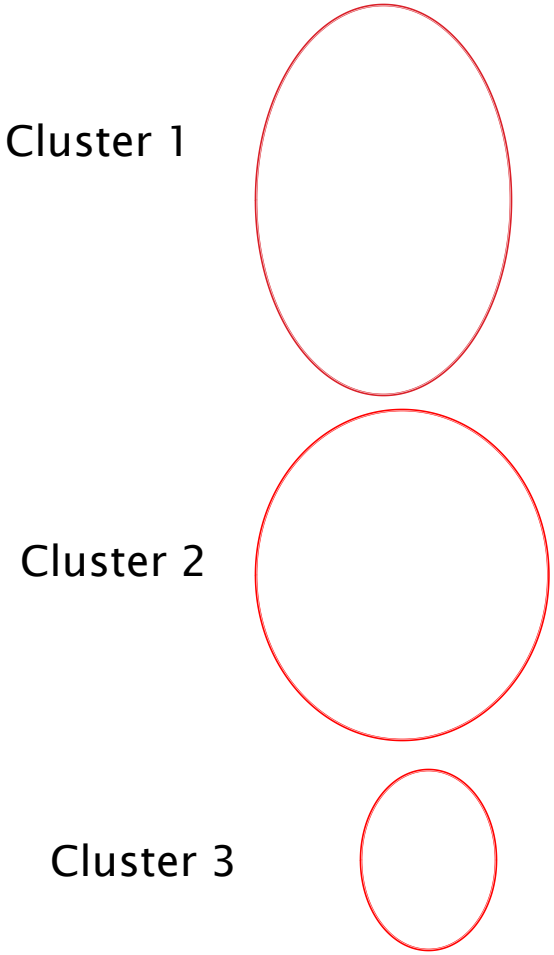
QUERY TERMS CLUSTERING ALGORITHM:

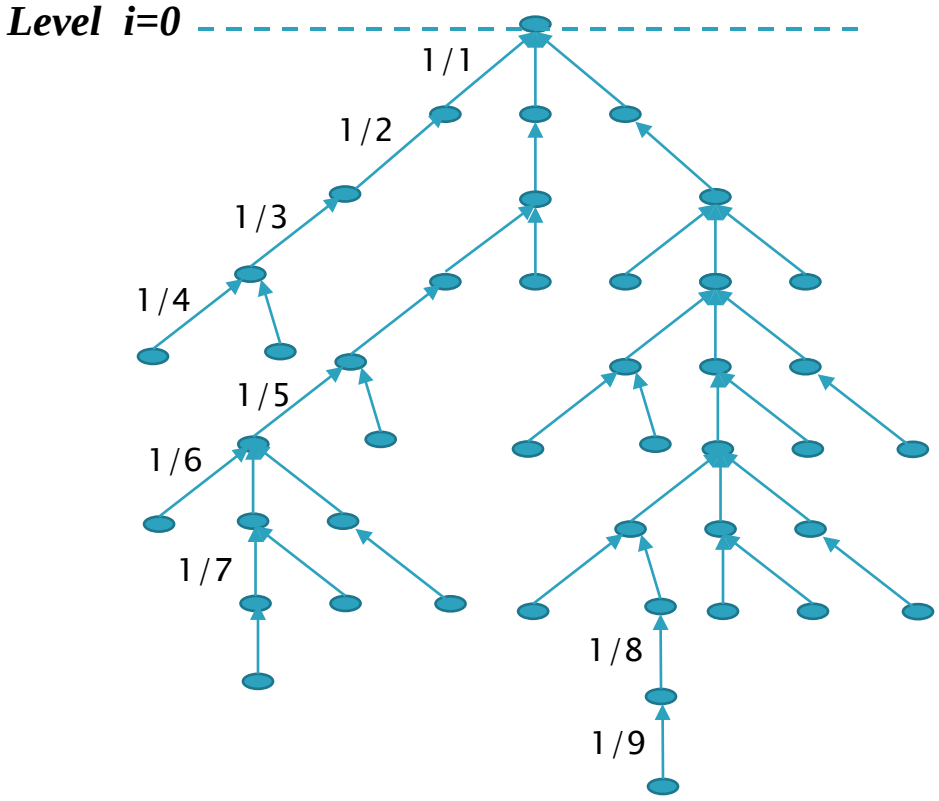
```
T    // Taxonomy with weighted links
E= {e0, e1...}  // set of query terms (nodes)
C = {}    // set of clusters
ci =      // ci  C
D     // distance function
ts = Value    // threshold
While Not (empty(E))
  e_d = deepest(E) // find the deepest term
  ci = ci  U {e_d }  // init. ci with the deepest term
  cluster_up(e_d , parentOf(e_d))
  C = C U {ci}
  E=E-{ci}
end
End
```

```
function cluster_up(predecessor, e)
If  D(e_d,e)• ts
  While has_children(e)
    if   childOf(e)• predecessor
    cluster_down(pull_childOf(e))
  end
  ci=ci U {e}
endif
cluster_up(e , parentOf(e))
End
```

```
function cluster_down(e)
If  D(e_d,e)• ts
  While (has_children(e))
    cluster_down(pull_childOf(e))
  end
  ci=ci U {e}
endif
end
```

Clustering Algorithm





Evaluation: test dataset

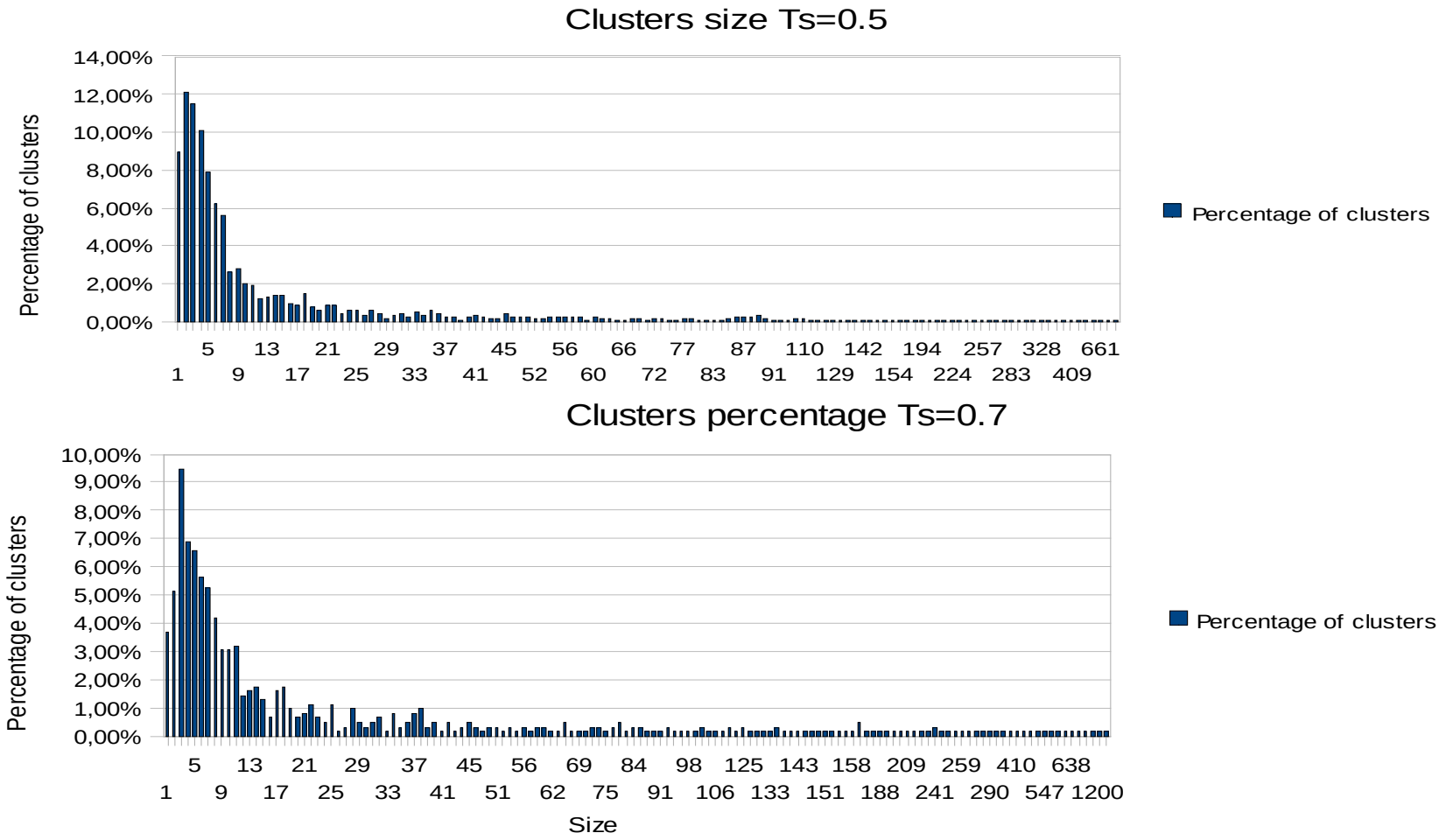
- AOL search logs
- 20 millions of queries collected over 650k users (USA) in a period of 3 months

AnonID	Query	QueryTime	ItemRank	ClickURL
2771158	california hospital association	19.03.2006 23:16		
2771158	glendale adventist medical center	19.03.2006 23:16	1	http://www.glendaleadventist.com
2771158	free electronic greeting card	20.03.2006 22:47		
2771158	csun webct	21.03.2006 08:01		
2771158	the bodega	22.03.2006 01:29		
2771158	the bodega pasadena	22.03.2006 01:29	1	http://losangeles.citysearch.com
2771158	the bodega pasadena	22.03.2006 01:29	2	http://www.pasadenacitycenter.com
2771158	el paseo mall pasadena	22.03.2006 01:35	2	http://www.englekirk.com
2771158	el paseo mall pasadena	22.03.2006 01:35	8	http://www.rubios.com
2771158	the bodega el paseo mall	22.03.2006 01:37		
2771158	the bodega el paseo mall	22.03.2006 01:37	13	http://www.apa.udel.edu
2771158	mapquest	22.03.2006 01:39	1	http://www.mapquest.com
2771158	hollywood fitness private trainers	22.03.2006 01:44		

Evaluation

- Objective cluster quality measures
- Manual study of cluster semantics
- Influence of threshold on cluster distribution

Experimentation (threshold tuning)



- The threshold is determined experimentally by tuning : it balances small clusters and too general clusters

Conclusion… Next step

- Efficient and fast user interests identification
- The threshold could be determined experimentally by tuning
- Clusters are inputs to the user communities discovery and resource aggregation processes
- Next…
  - Improvements/cluster quality evaluation
  - Users profiles/similarity (overlap), resource aggregation
  - Discover other potential applications in the “black box”

Thank you for your attention

Any questions ?

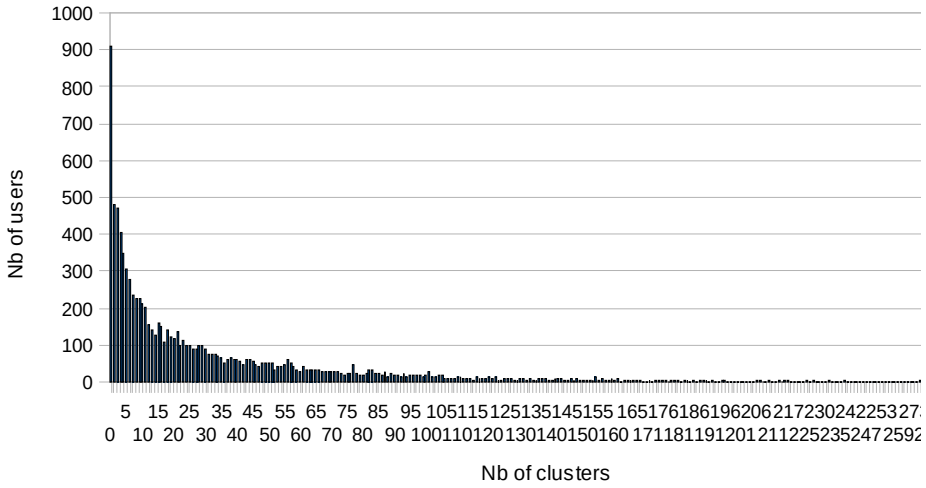
Users community and resource aggregation

- Depending on the adopted approach (global or local) the users grouping process is realized as :
  - Global: two users are considered to be in one group if they share the same clusters
  - Local: two users are in the same group if their corresponding clusters overlaps

Experimentation (users-clusters)

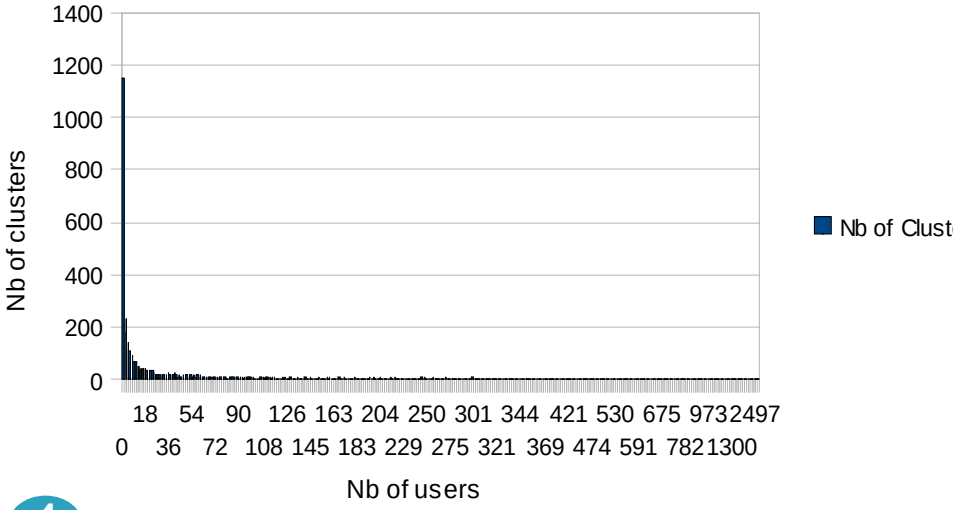
1

Users with the same Nb of clusters Ts=0.6



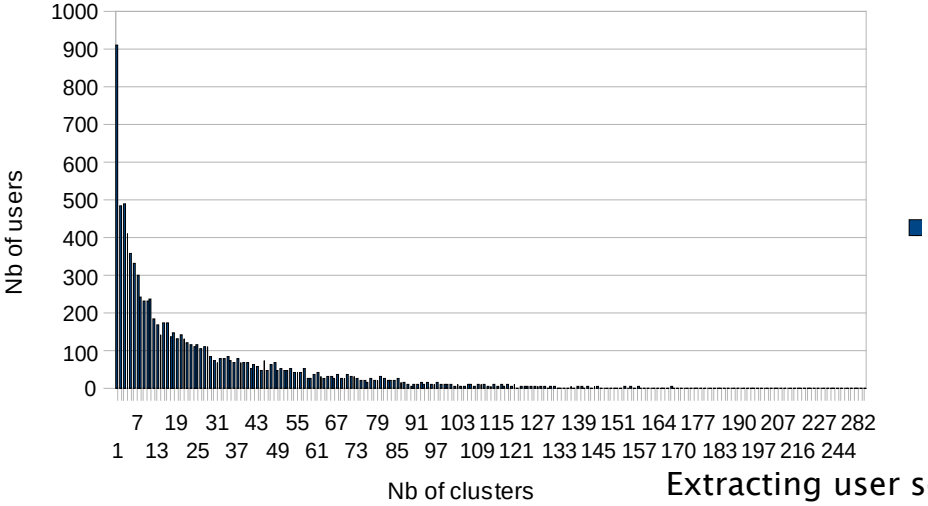
3

Clusters with the same Nb of users 0.6



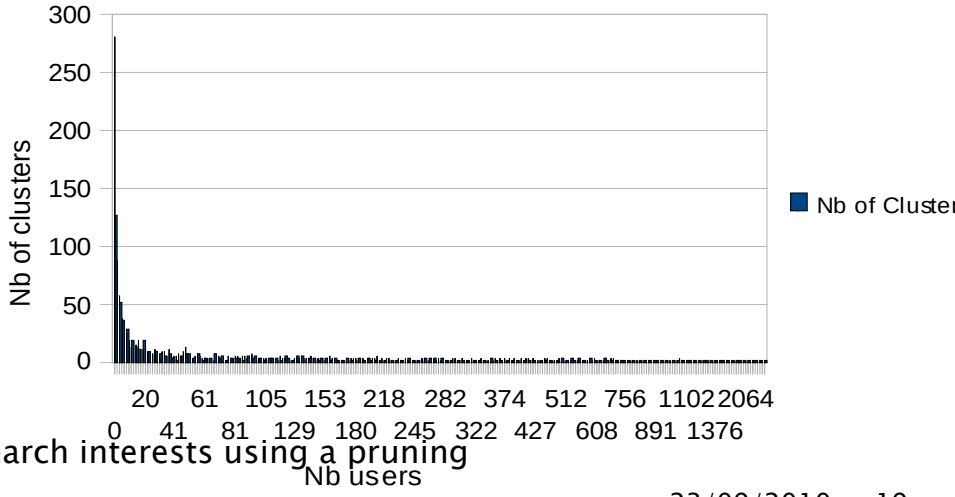
2

Users with the same Nb of clusters Ts=1.0



4

Clusters with the same Nb of users Ts=1.0



Outline

- Issue
- Framework for usage analysis
- Query terms clustering algorithm
- Experimentations
- Users community and resources aggregation
- Conclusion & Next step

Semantic distance function (proposal for improvement)

- In the context of clustering several improvement have been proposed:
  - Include the co-occurrence relationship in the distance function:

$$D'(x,y) = D(x,y) / C[x,y]$$

  - Include the terms frequency as it reflects the term importance

How to measure the efficiency of a distance/similarity measure ?

- Use of human judgment/similarity measure correlation proposed by Miller and Charles, the MC correlation
  - 30 pairs of nouns rated (0-4) by 38 native English speakers

Existing algorithms for clustering

- Hierarchical algorithms
  - Single linkage
  - Complete linkage
  - Average linkage
- Partitioning algorithms
  - K-means
- Graph algorithms
  - Neighborhood graph algorithm (spanning tree)
  - B-coloring