

# SEMANTIC COMPARISON OF KEYWORDS AGGREGATES

Christian Belbèze

*Université de Toulouse, Institut de Recherche en Informatique de Toulouse, UMR 5505,  
118 route de Narbonne,  
F-31062 Toulouse Cedex 9, France*

Max Chevalier

*Université de Toulouse, Institut de Recherche en Informatique de Toulouse, UMR 5505,  
118 route de Narbonne,  
F-31062 Toulouse Cedex 9, France*

*Université de Toulouse, Laboratoire de Gestion et de Cognition, EA 2043,  
129A, avenue de Rangueil - BP 67701,  
F-31077 Toulouse Cedex 4, France*

Chantal Soulé-Dupuy

*Université de Toulouse, Institut de Recherche en Informatique de Toulouse, UMR 5505,  
118 route de Narbonne,  
F-31062 Toulouse Cedex 9, France*

## ABSTRACT

In the context of information access in the social Web (Information Access 2.0), we present an original method to build automatically and dynamically users' communities based on past search queries. These "search communities" can for instance be used by any user to find relevant query terms or to find any other online user to discuss about his information needs. The proposed approach has been experimented in order to compare its results concerning particularly the semantic coherence of obtained communities. These experiments exploit the aol.com search engine log.

## KEYWORDS

Search community, query, graph, semantic coherence, user log file, search engine.

## 1. INTRODUCTION

To retrieve information relevant to their needs, most of users are commonly querying a search engine like Google or Yahoo! In this context, every user is alone facing the search engine without any real external support. With the raise of the participative web (Web 2.0), many solutions can be envisaged to introduce social approaches in the retrieval process. In this paper we focus on the construction of "search communities" where users can be automatically attached according to their needs for instance. These communities can be dynamically built in order to provide a more social dimension of retrieval process. For instance, allowing users to contact other people sharing his concerns in a transparent way, recommending additional query keywords to user's or defining search contexts are undoubtedly services desired by the users in order to obtain some support. A real implementation of such motivations can be seen through HAKIA<sup>1</sup> search engine (beta version) and its "meet others" functionality. These novel services can use our proposed construction method of communities to help users in a "social" manner.

In this paper, we present an original way to construct these "search communities" thanks to users' query terms clustering method. This method is evaluated in order for instance to measure semantic coherence of the constructed communities. Obtained results are discussed in this paper. This latter is organized as follows. Section 2 presents the general context and the state of the art of web communities and in a more specific way how queries can be clustered to build "search communities". Section 3 illustrates our proposal that is to say a query terms clustering method to build semantically coherent communities. Section 4 underlines the experiments done to evaluate this method.

---

<sup>1</sup> <http://www.hakia.com/>

## 2. CONTEXT AND STATE OF THE ART

Building communities is almost common on the context of the Web. For instance many works have been done around Web communities. Such communities are principally exploiting the hypertext structure of the web (thanks to HITS measure for instance) to build what we can call “content communities” (Balfe & Smyth, 2005; Zhang et al., 2006). We do not identify in the literature approaches which really exploit past queries to construct communities. Besides this, we have identified many works related to query clustering method that could be applied in order to build “search communities”.

Clustering keywords are not really applied to real search engine log files. The difficulty to get log files, including the keywords used in search engines, is explained by their commercial value and their direct impact in the E-market. Thus, there are few specific publications related to the study of relations between query terms in a real search engine.

Many works about the aggregation of keywords on specialized search engines can be found. Shingo Otsuka and Masaru Kitsuregawa (2006, pp. 842\_852) create clusters built by the association of the keywords having been historically used to select a page then validated by the fact that the words are contained in communities of sites (sites connected by bonds HTTP) associating the page. Other work as those done by Cui (2002, pp. 325-33 2) try to create semantic spaces of keywords by correlating past queries keywords with terms existing in any selected document (URL, title, keywords...) and sometimes as in work of Ohkubo, M., Sugizaki (1998, pp. 2250–2258) in a limited temporal space.

In our proposal we want to construct semantically coherent search communities based on past query terms. To do this, we used an adapted structure: the “clique” structure. Clique, in a graph, such as definite by Luce and Perry (1949), is a subset of a graph composed of three nodes at least that respects:

- every node composing the sub-graph must be adjacent (in relation) to all other nodes of the clique
- it does not exist any node outside the sub-graph which is in relation to all the nodes of clique

These structural constraints of a clique ensure to obtain a coherent semantic space. We will combine cliques in another type of regrouping technique based on the graph theory and works of C. Haussman (1997, pp.287\_299) and Christophe Jermann (2002). These methods introduce an extension operator which allows us to carry out aggregates by taking into account the importance of the number of joint uses of query keywords with the number of total use of the keyword itself.

As far as we know, there is no publication related to the semantic coherence of query keywords aggregates. As a solution, we propose a way to measure this aspect in this paper (section 4).

## 3. AN ORIGINAL CLUSTERING METHOD TO BUILD SEARCH COMMUNITIES

The aim of a clustering method applied to a given population, in order to create aggregates, is to minimize intra-cluster distance between objects and to maximize the inter-cluster distance. The relation (such as similarity) we use between objects (query terms) will be notified as  $A R B$  meaning that the object (i.e. keyword) “A” is used jointly with the object (i.e. keyword) “B” in at least one case (query).

### 3.1 Clustering keywords with an implementation the Hoffmann, Lomonosov and Sitharam’s method (HLS)

To obtain bigger but still coherent aggregates we use Hoffmann, Lomonosov and Sitharam (HLS) method which provide an important flexibility to the structure. It relies on organized triad structures (cliques composed of 3 nodes).

#### 3.1.1 Definition of HLS

This method, conceptually suggested initially by Hoffman, Lomonosov and Sitharam (Hoffmann and Al, 1997; 1998; 2000), is a method of ascending structural decomposition. It aims at finding “rigid” aggregates of objects. These aggregates are then assembled recursively. The method is an example of the recursive methods of rigidification of **GCSP** (Geometric Constraint Satisfaction Problem).

**Analyzing phases of HLS method:** A first phase of analysis consists in gathering the aggregates as long as regroupings are possible. The phase of analysis is in three parts: fusion, extension and condensation. The **Fusion** phase seeks the aggregates of minimal size. The **extension** phase consists in including a nearby object in the current aggregate and this, as long as a nearby object to insert exists, using an extension operator. The **Condensation** updates the plan of assembly.

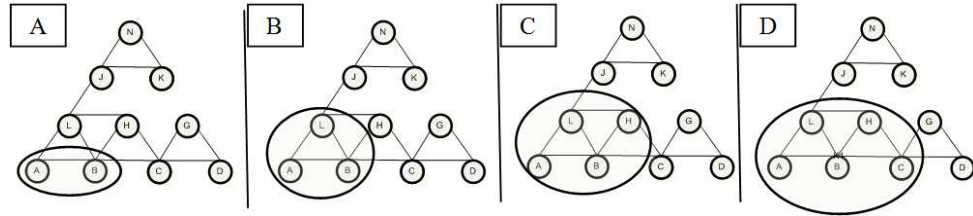
### 3.1.2 Implementation of the HLS's conceptual method

In our study we will define the minimum aggregate as one clique. The phase of fusion will thus seek these structures. In the original proposal, Hoffman purposes to use the method on a non directed graph and do not give any suggestion about how to build the operator of extension. In our implementation we use a weighted directed graph and we defined the operation of extension using an operator in respect of the following characteristics:

**Operator of extension "O":** the graph related to an aggregate must always remain "biconnex" during extensions.

**"Biconnex" graph:** a "biconnex" graph is a graph where each point is connected by at least two ways towards any other point of the graph (cf. figure 1).

Figure 1 - Illustration of the unfolding of the algorithm Fusion/Extension



**Weight:** We will name "*Weight*" the number of researches related to an object. The weight of a keyword is thus the number of requests including this keyword. The weight of a **R<sub>ab</sub>** relation has being defined as the number of requests including the two keywords A and B.

**The keyword weight:** considering A a keyword, N<sub>b</sub> being the total number of queries, function MC<sub>iA</sub> returns 1 if the A exists in the i<sup>th</sup> query, 0 if the A does not occurs in i<sup>th</sup> query. The weight of a keyword A noted W<sub>A</sub>, is calculated by:

$$W_A = \sum_{I=1}^{N_b} MC_{iA}$$

**The weight of a relation:** considering a word A and a word B, if a R<sub>AB</sub> relation exists such as A R B, N<sub>b</sub> the total number of queries, function IH<sub>RAB</sub> returns 1 if the keywords A and B are present in at least in one query, 0 if A and B does not occurs in any query. The weight of a RAB relation noted W<sub>RAB</sub> is calculated by:

$$W_{RAB} = \sum_{I=1}^{N_b} IH_{RAB}$$

Use of the weight for orienting the graph and to improve the extension operator: We will thus improve the extension operator by taking into account the concept of weight in a graph. Orientating the graph is done thanks to the ratio CRB (Coefficient of Reliability of Bond) of the weight of the relation on the weight of the keyword (cf. Table 1 - Matrix asymmetrical Figure 2). This ratio will be computed for a word A in relation with the word B noted R<sub>AB</sub>, for W<sub>A</sub> the weight of the word A, and for W<sub>RAB</sub> the weight of the R<sub>AB</sub> relation. The Coefficient of Reliability of Bond (CRB) of the word A towards the word B noted CRB<sub>A=>B</sub> is calculated by:

$$CRB_{A \Rightarrow B} = \frac{W_A}{W_{R_{AB}}}$$

Figure 2 - Graph directed of the CRB of the matrix presented of Table 1 - Matrix asymmetrical (CRB is presented in % to facilitate its interpretation)

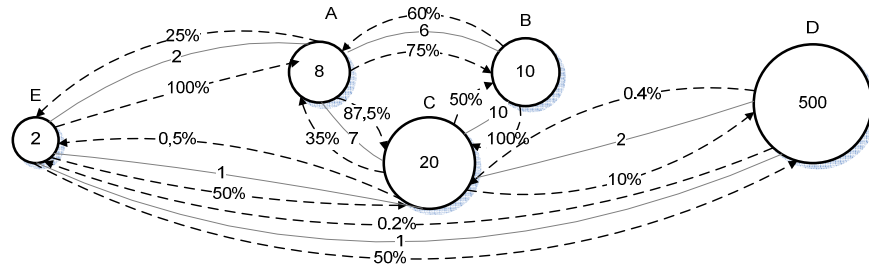
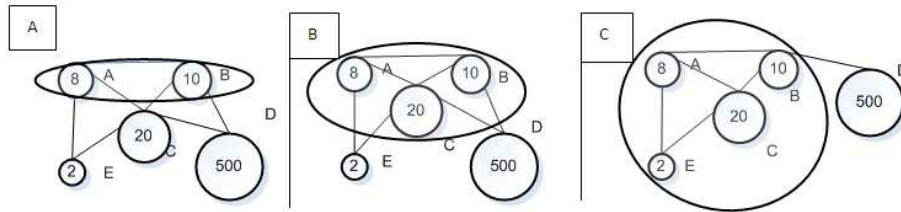


Table 1 - symmetrical and asymmetrical weights matrices of non directed and directed graphs

Symmetrical undirected graph matrix						Asymmetrical directed graph matrix - CRB							
Word	Weight	A	B	C	D	E	Word	$R_{CRB}$	A	B	C	D	E
A	8	-	6	7	0	2	A	->	-	.75	.875	0	.25
B	10	6	-	10	0	0	B	->	.60	-	1	0	0
C	20	7	10	-	2	1	C	->	.35	.5	-	.1	.05
D	500	0	0	2	-	1	D	->	0	0	.004	-	.002
E	2	2	0	1	1	-	E	->	1	0	.5	.5	-

**The extension operator:** using this algorithm without an operator of extension had as a consequence the creation of massive aggregates of several thousands of keywords. As a conclusion we define thresholds of validity. In order to prune the graph we are deleting relations having a CRB value lower than a specific threshold. We only keep relations having a CRB higher than a prevailed value called Minimal Value of CRB or MVCRB. In order to keep low-weight words in the structure, we will maintain, whatever the CRB of opposite direction, all relations having a CRB higher than the Value of Activation or VACRB (cf. Figure 3).

Figure 3 - Unfolding of the algorithm Fusion/Extension using the operator of extension



The final extension operator must respect:

- the graph must remain "biconnex"
- a relation having a CRB lower than MVCRB should be removed except if the CRB of opposite relation is higher to VACLF

In figure 3, according to the Hoffmann method, with MVCRB=5 and VACRB = 20, the relation between C and D should not be maintained because the  $CRB_{D \rightarrow C}$  is lower than the fixed MVCRB and the  $CRB_{C \rightarrow D}$  is lower than the fixed VACRB. The element "D" cannot thus join the aggregate because the resulting graph would not be then "biconnex" any more.

In respect of these elements we defined the following algorithm (algorithm 1) which presents different steps of the HLS method.

Algorithm 1 - Application of method HLS

```

For each keyword X [Phase of fusion]
  To extract keywords Y[] which form a valid triad with X
  For each couple of valid keywords X-Y[] [Phase of extension]
    If there is not any aggregate containing couple X-Y[] and that the couple was not tested?
      Create a new aggregate "X there" and addition of X there
      As long as add keywords in the aggregate
        For the words of the aggregate
          Scan for new words in triad
          Add keywords forming a triad with the words of the aggregate
          Notation of the couples found like "tested"
        End of For
      End of As long as
    End of If
  End of For [End of Phase of extension]
End of For [End of Phase of Fusion]
  
```

## 3.2 Postulate and proposal for a semantic validation technique

### 3.2.1 Postulate: two laws

- 1 - Internet mainly consists of Web sites and semantically coherent documents.
- 2 - Internet search engines users are aware and have sufficient experience to use keywords having linked together and with the required subject.

### 3.2.2 Proposal for a semantic validation technique

To evaluate the semantic coherence of obtained communities, we propose to compare the number of sites retrieved by users' queries with queries randomly built. For instance, table 2 shows the number of retrieved document resulting from three user's queries (1, 2, 3) composed of three keywords and a fourth one (4) randomly composed of keywords coming from initial queries. To obtain these results we use the aol.com search engines.

Table 2 - Method of semantic validation: Acquisition of the value for the triads

ID	Queries	Nb Sites	ID	Queries	Nb Sites
1	Einstein relativity equation	<b>230 000</b>	3	jazz saxophone selmer	<b>115 000</b>
2	cream spinach butters	<b>413 000</b>	4	Einstein spinach selmer	<b>141</b>

It is easy to understand that the last query (4) has no sense and thus few documents are relevant to this query. So the other queries can be qualified as "semantically coherent" since they retrieve more documents than a query randomly built.

### 3.3.3 Method of compared semantic validation

We do not claim to describe here a method of absolute semantic validation, but rather an indicator which will make it possible to carry out comparisons between methods of regrouping and their evolutions. We propose the comparison of the number of sites, found by the search engine of the aol.com site, between three set of triads. The first set is constituted by triad (three terms coming from real queries) used by a real user, second set contains combinations of three keywords extracted from the aggregates and the third set contains randomly combined three keywords queries (baseline). All the combinations of three words or triad of each aggregate will be presented to the search engine. That represents **792756** (seven hundred and eighty-two thousand seven hundred and fifty-six) combinations for the aggregates of the method on the sample of study (see next section). The random sample was made of **500000** (five hundred thousands) triads of different words taken randomly on the sample.

## 4. EXPERIMENTATION

### 4.1 Creation of a sample

#### 4.1.1 The material

We supported our work on an extract aol.com search engine log files. This extract integrates thirty three million queries carried out from March 1<sup>st</sup>, 2006 to April 30<sup>th</sup>, 2006. These queries are mainly in English. The structure of the file integrates an identifier, the date and the hour of the query.... This file is available on the <http://gregsadetsky.com/aol-data> site for study purposes.

#### 4.1.2 Preparation of the sample

In order to work on a representative and nevertheless easy to handle sample we made the choice to limit our experiments to queries from only one day: april 17<sup>th</sup>, 2006. We applied several rules to treat these queries:

- the keywords are defined like a set of letters without space. any space character is thus interpreted as a keyword separator
- the quotation marks ("") as all the elements of punctuation were replaced by white spaces
- words having a length equal to one character were deleted. Many words considered to be not significant (like stop words) were also deleted from the study
- only queries having two words and more were used in our study

We then drew aside from the study a list of words (75) considered as vacuums (the, this, I, we ...).

In order to avoid handling words with no made sense because of a great use we decided not to consider the words having been used in more than 1000 (thousand) queries. To draw aside these words which are by definition the least discriminating to us, makes allow us to hope avoiding the construction of mega-aggregates centered on these keywords. The full number of queries studied in the sample of april 17<sup>th</sup>, 2006 being of more than 200000. These words are 14 out of 51994 studied keywords thus 0.027% of the sample.

**The final sample:** once these various “filters” applied: the object of the study is a set of: **51980 keywords used in 200646 queries.**

### 4.3 Result of the clustering by the HLS-CRB method

**MVCRB and VACRB or minimal Value activation of CRB:** After several tests on samples, we defined threshold values: Minimal Value of CRB or **MVCRB** with **5%** of the weight of the word and the Value of Activation or **VACRB** at **20%** of the weight of the word. These values could be modified or adjusted during next experimentations. They are only given as an example and their value is out of the scope of this study.

**Results:** The final results are from **24537** different words a set of **9556** aggregates. The median number of keywords per aggregate is **4.04**. The most important aggregate include **133** keywords.

### 4.4 Result and analyze of the compared semantic validation method

#### 4.4.1 Chart of the semantic validation test

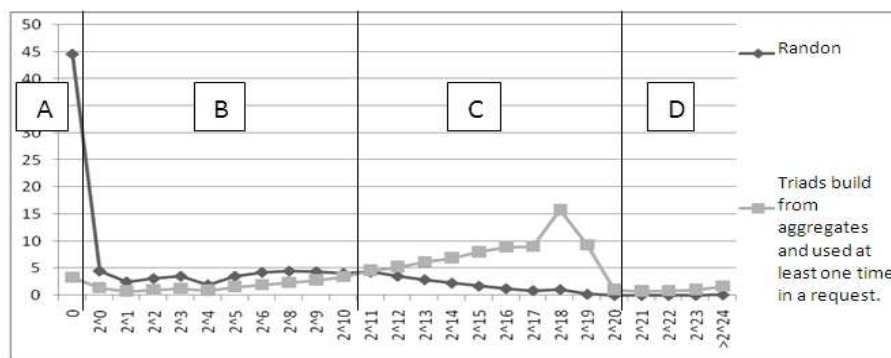
We will represent the results in a semi-logarithmic scale. A semi-logarithmic reference trend graph is reference trend graph which one of the axes. Here that of the ordinates, is graduated according to a linear scale and that of the X-coordinates (X), is graduated according to a logarithmic scale. The advantage of a semi-logarithmic representation is its aptitude to represent measurements which are spread out over extremely broad values. Semi-logarithmic representations in power of 2 were already used by Zipf (1935).

- **Ordinate:** We will place in ordinate the percentage of combinations found by class.
- **X-coordinates:** We gathered the number of sites turned over in classes and these classes in a logarithmic scale. In order to remain on the finest possible classes we chose classes by power of 2.

#### 4.4.2 Search of a comparative data

With a view to highlight a particular zone and differences we compare here the two curves response of the two most distant spaces semantically.

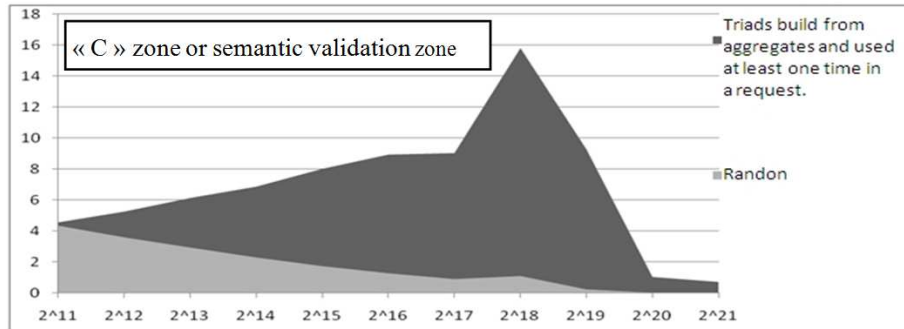
Figure 4 - Comparison of the two curves most distant semantically according to the postulate from paragraph 3.3.1 and determination of the remarkable zones



We will thus compare the curve resulting from the words combined by chance with the curve of reference resulting from the test of triads for which there is at least a research including these three words and drawn from aggregates created with the method of HLS-CRB. Except for the random triads, the others triads tested are extracted from aggregates obtained by the HLS-CRB method. The comparison is first carried out here in a graphic manner. We notice four clearly identifiable zones (cf. Figure 4): the zone A of the 0, the zone B of  $2^0$  with  $2^{11}$ , the zone C of  $2^{12}$  with  $2^{21}$  and the zone D higher than  $2^{21}$ .

The zones “B” and “D” do not represent much interest, the curves not presenting a notable difference. The zone “C” is the most different zone. The differences noted on the zone “C” are compensated in the zone “A”. The zone “A” is reduced to only one value “0” and is too narrow to be exploited. In order to better perceive the importance of “C”, let us read again the graph by omitting the zones A, B and D (Figure 5).

Figure 5 - Zoom on the Zone “C” selected like zone of study



#### 4.4.3 Calculation of the SCCV (Semantic Coefficient of Compared Validation):

We will compare the surface taken by the two histograms. The SCCV or semantic coefficient of compared validation, having then the value “1” for the equivalence of the histogram of the triads of three words having been at least once used in the same research and 0 for the value of the histogram of the random triads. The mathematical formula will thus be for a particular curve to validate Xa: Let us note **Ar** the surface of the histogram of the triads of which all the words are included at least once all together in a research, **Aa** the value of the surface of the random triads histogram and **Ax** value of the surface of the triads histogram to be compared

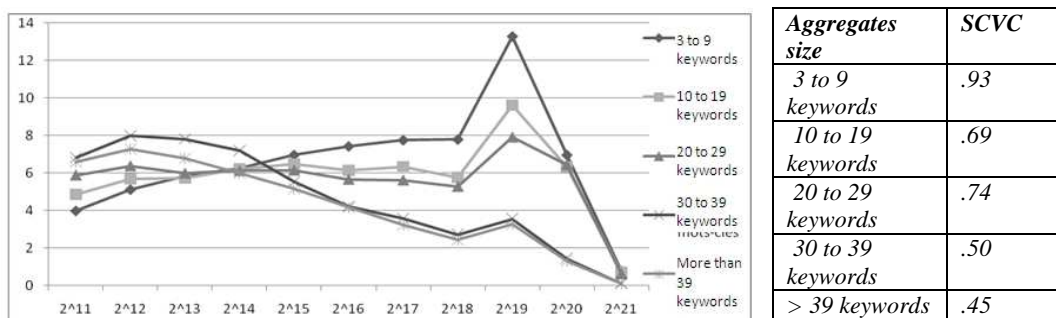
$$Ar = \sum_{i=11}^{21} Yi = 75.4 \quad Aa = \sum_{i=11}^{21} Y'i = 22.8 \quad Ax = \sum_{i=11}^{21} Y''i$$

$$\text{Let us note for a curve X, } SCCV_x = \frac{(Ax - Aa)}{Ar - Aa}$$

#### 4.4.4 Comparison of coefficients SCVC (Semantic Coefficient of Validation Compared) for aggregates of different size.

In order to determine a target for future work it seems important to limit on the top the size of the aggregates. We will compare here the SCVC for aggregates of different size. We notice either graphically (cf. Figure 6) or by the calculation of the SCVC that the more the number of keywords is important the more SCVC tends to drop.

Figure 6 - Comparison of the SCVC according to the size of the aggregates in Zone “C”



The rupture between the aggregates of size lower than 30 keywords (cf. Figure 6) and that higher or equalizes is clearly marked. If one places an acceptable level of SCVC at a value .5 that allows us to affirm that statistically the aggregates of more than 30 keywords do not present a good SCVC.

## 5. CONCLUSION

To give new services to users we need to profile and to regroup users. Clusters of keywords could be use as connection points between users themselves and between users and services.

The search for a semantic measurement of keywords regrouping system by using Internet as a validator passes without any doubt through comparative method.

The HLS-CRB method of regrouping can evolve/move and give results different according to the parameters from CRB. Method of semantic compared validation allows arbitrating the quality of these evolutions. Other methods of regrouping can also be compared.

In this first approach we regarded the keywords as neutral and independent objects. In the future we can improve the method by using the Porter algorithm, WorldNet integration, other ontological dictionaries, synonyms dictionary or word communities build from dictionary as this is purpose by Bruno Gaume (2008). The reduction of large aggregates higher than 30 keywords could also be done by a modification of index CRB (Coefficient of Reliability of Bond) and a setting in quarantine better controlled on-used or empty keywords. At least, we can compare aggregates characteristics (clustering coefficient, average distance and diameter) with results returned by the semantic comparison method.

## 6. REFERENCES

- Balfe, E. and Smyth, B., 2005. A Comparative Analysis of Query Similarity Metrics for Community-Based Web Search. *ICCBR 2005*, H. Munoz-Avila and F. Ricci (Eds.), LNAI 3620, pp. 63–77.
- Bruno, G. and Fabien, M., 2008, From Random Graph to Small Word by Wandering, eprint arXiv:0804.0149.
- Christophe, J., 2002, Résolution de contraintes géométriques par rigidation récursive et propagation d'intervalles, *Journal Électronique d'Intelligence Artificielle*, <http://jedai.afia-france.org/>
- Christophe, J., 2003, Résolution de contraintes géométriques par rigidification récursive et propagation d'intervalles, *Laboratoire d'Informatique de Nantes-Atlantique*. pp. 104, pp.121-160
- Cui, H., Wen et al 2002, Probabilistic query expansion using query logs. *Proceedings of the eleventh international conference on World Wide Web*, pp. 325 – 332.
- Festinger, L., 1949, The analysis of sociograms using matrix algebra. *Human Relations*, 2, 153–158.
- Fonseca, B.M. et al, 2003 Using association rules to discover search engines related queries. *First Latin American Web Congress (LAWEB'03)*, pp. 66-71.
- Fu, L. et al , 2003, Collaborative querying through a hybrid query clustering approach, 2003, *Digital libraries: Technology and management of indigenous knowledge for global access*, ICADL, pp. 111-122.
- Fu, L. et al, 2004, Collaborative querying for enhanced information retrieval, 2004 *European conference on research and advanced technology for digital libraries*, pp. 378-388.
- Gangnet, M. and Rosenberg, B., 1993, Constraint programming and graph algorithms, *Annals of Mathematics and Artificial Intelligence* 8(3–4): 271–284.
- Hoffman, C. and Jaon-Arinyo, 1997. Symbolic constraints in constructive geometric constraint solving. *Journal of Symbolic Computation*, 23:287-299.
- Hoffmann, C. et al, 2000, Decomposition plans for geometric constraint systems, *Proc. J. Symbolic Computation 2000*.
- Koutsoupias, 2000, N.: Exploring web access logs with correspondence analysis. *Methods Bransford, J.D., Brown, A.L.O., & Cocking, R.R. (Eds.)*.
- Latapy, M., 2007. *Grands graphes de terrain – mesure et métrologie, analyse, modélisation, algorithmique. Habilitation à diriger des recherches*, Université Pierre et Marie Curie, Paris, France.
- Luce, R.D. and Perry, A.D. , 1949, A Method of matrix analysis of group structure, *Psychometrika*. 14, pp. 95-116
- Ohkubo, M. et al, 1998. Extracting information demand by analyzing a www search log. *IPSJ Journal* 39(7) (1998) pp. 2250–2258
- Reinhard, D., 2005, *Graph Theory Electronic*, Springer-Verlag Heidelberg,
- Shingo, O and Masaru, K., 2006. Clustering of Search Engine Keywords Using Access Logs, *Database and Expert Systems Applications*, Springer Berlin / Heidelberg, pp. 842-852
- Zipf, G. K., 1935, *The Psychobiology of Language, an Introduction to Dynamic Philology*, Boston, Houghton-Mifflin
- Zhang, Y. et al., 2006. *Web Communities, analysis and construction*. Springer-Verlag Berlin Heidelberg. ISBN 3-540-27737-4.