
Nouveaux services basés sur des agrégations de mots-clés validées sémantiquement

Christian Belbèze* — Max Chevalier^{*,**} — Chantal Soulé-Dupuy*

* Université de Toulouse, Institut de Recherche en Informatique de Toulouse, UMR 5505,

118 route de Narbonne,
F-31062 Toulouse Cedex

** Université de Toulouse, Laboratoire de Gestion et de Cognition, EA 2043

129A, avenue de Rangueil - BP 67701

F-31077 TOULOUSE Cedex 4

christian@belbeze.com, Max.Chevalier@irit.fr, chantal.soule-dupuy@univ-tlse1.fr

RÉSUMÉ. *A l'heure du Web social, nous présentons une solution pouvant servir de base à de nouveaux services tels que la construction automatique et dynamique de communautés d'utilisateurs. Pour cela, nous proposons de créer des agrégats de mots clés issus des recherches antérieures des utilisateurs pour construire ces communautés. Nous présentons la démarche que nous avons suivie pour obtenir un algorithme de regroupement des mots-clés provenant de fichiers de traçage (log) ; nous illustrons cet algorithme au travers de son application au fichier de traçage du moteur de recherche aol.com. Afin d'évaluer les résultats obtenus, nous proposons une démarche permettant, par comparaison, l'obtention d'un coefficient de cohérence sémantique des agrégats ainsi créés. Dans une expérimentation nous mesurerons la perte de cohérence sémantique liée à l'augmentation de la taille des agrégats. L'intérêt de l'approche proposée réside dans le fait qu'elle peut être exploitée pour offrir encore plus de services à l'utilisateur.*

ABSTRACT. *At the time of social Web, we present a solution able to build new services offered to users such as the automatic and dynamic construction of users' communities. These communities are done by the creation of semantically coherent aggregates, built from keywords from previous users' queries. We present algorithms for keywords gathering; these keywords are resulting from aol.com search engine log files. To evaluate these results, we developed a methodology allowing, by comparison, to obtain a semantic coherence coefficient on the aggregates thus created. As an experiment we measure the loss of semantic coherence related to the increase of the aggregates size. The main advantage of the proposed approach is that it can be exploited to offer the user more services*

MOTS-CLÉS : *Mot-clé, agrégat, cluster, sémantique, graphe, fichier de traçage, log, moteur de recherche, groupe, sac, Internet.*

KEYWORDS: *Keywords, aggregate, cluster, semantic, graph, log files, search engine, group, bag, Internet.*

1. Introduction

Pouvoir mettre en contact de manière transparente un utilisateur avec une communauté partageant ses préoccupations, proposer des mots-clés supplémentaires dans une recherche ou définir des contextes de recherche sont des services susceptibles d'aider les utilisateurs des moteurs de recherche sur Internet à accéder à toute information utile, voire à optimiser l'accès à cette information par un partage implicite de compétences. On parlera alors de création dynamique de communautés d'utilisateurs. Parmi les différentes méthodes qui pourraient être envisagées pour créer ou identifier ces communautés, nous avons choisi de nous intéresser à celles basées sur la création d'agrégats de mots-clés sémantiquement cohérents. La cohérence sémantique est une notion que nous définirons comme la capacité d'un groupe de mots à définir un champs d'un domaine le plus précis possible. Ces agrégats de mots-clés mis en correspondance avec les mots-clés d'une recherche d'un utilisateur permettront par exemple de rapprocher cet utilisateur des utilisateurs attachés aux agrégats les plus proches et ainsi de lui offrir de nouveaux services.

Dans cet article, nous présentons une méthode complète de regroupement de mots-clés en agrégats sémantiquement homogènes. Nous nous sommes orientés vers une approche de résolution de contraintes à base de graphes. L'approche étudiée repose sur des principes énoncés dans la méthode proposée par Hoffmann, Lomonosov et Sitharam en 1997, appelée « méthode HLS » (Hoffmann, Lomonosov et Sitharam, 1997). Nous proposons en particulier une modification appropriée de l'opérateur d'extension et des algorithmes de regroupements de mots-clés. Une technique d'évaluation a été également proposée afin de vérifier la validité des résultats obtenus. Cette évaluation de la méthode proposée a été effectuée sur un espace de mots-clés correspondant à un extrait d'une journée de logs du moteur de recherche d'aol.com. Les regroupements sont basés sur l'utilisation conjointe des mots-clés par un usager du moteur de recherche. Enfin, des comparaisons entre les réponses du moteur de recherche des agrégats de mots issus d'ensembles créés aléatoirement ou par les techniques de regroupement nous permettront de mieux définir les limites du système proposé et d'y apporter des améliorations.

La suite de cet article est organisée de la manière suivante : la section 2 est consacrée à un état de l'art sur les approches pour la création d'agrégats de mots-clés. La section 3 présente l'approche proposée, notamment les adaptations effectuées à partir de la méthode HLS, et la technique de validation sémantique associée. La section 4 illustre la démarche par la description d'une expérimentation à partir d'un fichier de logs d'aol.com et des résultats obtenus. Enfin, la section 5 conclut l'article par un bilan et les perspectives d'évolution de l'approche proposée.

2. Contexte et état de l'art

L'agrégation de mots-clés a fait l'objet de nombreux travaux ces dernières années tant en classification (de documents, de requêtes, de sites web, etc...) qu'en

recherche d'information. Or comme l'ont souligné d'autres auteurs avant nous (Shingo et al, 2006), l'étude des mots-clés utilisés dans le cadre des activités de requête des internautes via les moteurs de recherche « commerciaux » (Google, Yahoo, Exalead...) est difficile, voire quasiment impossible, du simple fait que les ressources nécessaires ne sont pas diffusées car elles représentent une partie de leur fond de commerce (exemple : revente des mots-clés). Il y a de fait peu de publications disponibles sur l'étude voire l'exploitation que l'on peut proposer des mots-clés utilisés dans les moteurs de recherche sur Internet. Nous allons toutefois dresser un état de l'art des travaux qui se sont intéressés à l'agrégation de mots-clés. Dans un premier temps nous discuterons des travaux s'intéressant aux regroupements de mots clés dans divers cadres de requête sur Internet. Par la suite, nous nous focaliserons sur les travaux qui se sont intéressés à la création d'agrégats sémantiquement homogènes qui ont inspirés nos travaux.

Certains travaux réalisés sur l'agrégation de mots-clés depuis des moteurs de recherche spécialisés s'appuient sur le contenu des sites Web sélectionnés par l'internaute au cours de sa recherche. En particulier, O. Shingo *et al.* (Shingo et Masaru, 2006) proposent la création de clusters construits par l'association des mots-clés ayant historiquement servi à sélectionner une page ensuite validée par le fait que les mots sont contenus dans des communautés de sites (sites reliés par des liens http) incluant la page sélectionnée. D'autres travaux, comme ceux de H. Cui *et al.* (Cui *et al.*, 2002) et de B.M. Fonseca *et al.* (Fonseca *et al.*, 2004) tentent de créer des espaces sémantiques de mots-clés en corrélant les mots-clés utilisés dans la recherche avec ceux mis en avant par les URL retournés (URL, titre, keywords, ...) et sélectionnés ensuite par l'internaute et parfois, comme dans les travaux de Cui et al., dans une fenêtre temporelle limitée. Les mots-clés sont alors placés dans un même cluster s'ils ont amené l'internaute à cliquer sur la même page ou le même réseau de pages. Les travaux de N. Koutsoupas (Koutsoupas, 2000) quant à eux ont pour but de créer une technique d'enrichissement de la requête en proposant à l'internaute un complément à ses mots-clés.

La création d'agrégats de mots-clés sémantiquement homogènes a également fait l'objet d'un certain nombre de travaux et semble plus appropriée à la proposition de nouveaux services aux internautes, notamment pour la définition de communautés d'usages. L'action d'« utilisation conjointe » est dans ce cas considérée comme élément de liaison. Il existe plusieurs méthodes de regroupements d'objets. Une première catégorie de méthodes s'intéresse aux structures appelées « cliques ». La définition de la *clique* a été originellement posée par L. Festinger (Festinger, 1949) ainsi que R.D. Luce *et al.* (Luce et Perry, 1949). Telle que définie par R.D. Luce et al. (Luce et Perry, 1949), la clique dans un graphe est un sous ensemble d'un graphe d'au moins trois nœuds, où chaque nœud est adjacent (en relation) avec tous les autres nœuds de la clique et tel qu'il n'existe pas d'autre nœud en relation avec tous les autres nœuds de la clique. Dans une clique, chaque nœud est donc en relation avec tous les autres nœuds de l'ensemble. Cette caractéristique semble être un point essentiel dans la constitution d'un espace sémantique cohérent. Une deuxième

catégorie de méthodes de regroupement de mots-clés repose principalement sur des approches à base de graphes pour la résolution de contraintes. Les travaux de Hoffman *et al.* (Hoffman *et al.*, 1997) et de Jermann *et al.* (Jermann, 2002) entrent dans cette catégorie. Les méthodes qu'ils proposent présentent la particularité de pouvoir utiliser un opérateur d'extension qui permet d'effectuer des regroupements en tenant compte de l'importance du nombre d'utilisations conjointes des mots-clés relativement au nombre total d'utilisations du mot-clé lui-même. Le caractère général de ces méthodes offre de nombreuses possibilités et notamment l'adaptation de cet opérateur d'extension pour la définition de nouvelles contraintes sur les utilisations conjointes entre mots-clés trop rares pour être représentatives.

En conclusion de cet état de l'art, il est important de noter que l'étude des méthodes de regroupements de mots-clés passe par l'analyse d'algorithmes proposés dans différents domaines de l'informatique et que ceux qui nous ont semblé les plus prometteurs reposent sur des approches de résolution de contraintes à base de graphes. Il n'en demeure pas moins que la recherche sur la validation sémantique d'agrégats de mots demeure un champ d'investigation encore largement ouvert.

3. Vers la proposition de nouveaux services aux utilisateurs de moteur de recherche sur Internet basés sur des agrégats de mots-clés

3.1 Propositions de méthodes de regroupements de mots-clés

L'utilisation d'Internet et des moteurs de recherche se fait aujourd'hui majoritairement de manière anonyme. Les seules informations connues alors sur l'internaute pendant sa recherche, hormis son matériel et ses logiciels, sont sa localisation réseau et les mots-clés utilisés dans ses recherches ainsi que les liens sélectionnés. Les internautes rechignent à faire un effort d'authentification et d'auto description. Les efforts d'authentification sont d'autant plus mal acceptés qu'ils ne correspondent, le plus souvent, qu'à un espace ou un usage réduit. Le temps de l'auto-description lui aussi, ne correspond pas aux habitudes d'immédiateté des services les plus consommés tels que les moteurs de recherche.

La création de communautés dynamiques permettrait à un utilisateur de coopérer avec d'autres sans avoir ni à s'authentifier, ni à se décrire, ni même à s'inscrire dans ces espaces. Il ne reste pas moins que la signature ou un élément de communication permanent, comme une adresse de messagerie électronique, permettront un fonctionnement asynchrone du système.

Dans les figures 1 et 2, Georges recherche, sans être connu par le système, des sites utilisant les mots-clés B, E, G et H. Grâce à l'agrégat N°1 contenant ces mots, Georges se voit proposer de rentrer en contact avec des utilisateurs ayant des centres d'intérêts proches des siens. Il peut soit ouvrir un salon de discussion pour qu'y soient invités automatiquement des Internautes concernés par les mots-clés de l'agrégat numéro1, soit démarrer une conversation en messagerie instantanée avec l'utilisateur « Anonyme » ou bien laisser un message à Annie (cf. Figure 2).

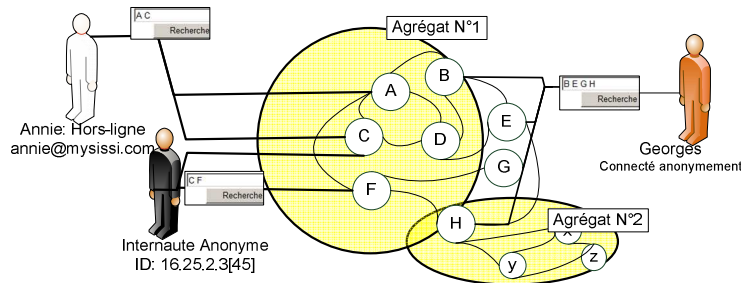


Figure 1. Attachement des Internauteurs à un agrégat en fonction de leurs recherches.

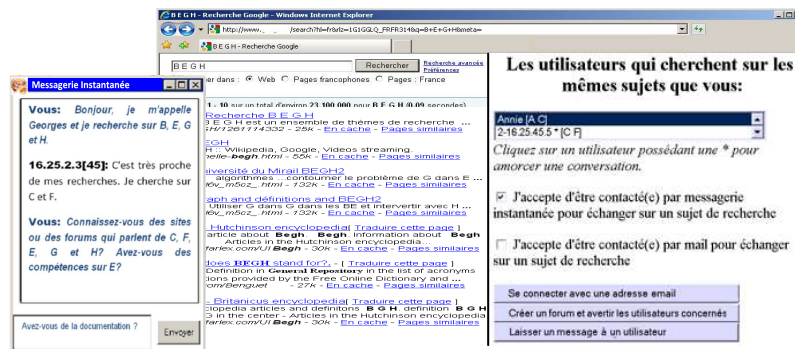


Figure 2. Utilisation des nouveaux services de communication au sein de la communauté dynamique.

L'objectif des travaux que nous présentons dans cet article est alors de définir une méthode de création de ces agrégats de mots-clés auxquels un utilisateur pourra s'identifier. De fait, le regroupement ou la création d'agrégats a, dans une population donnée, pour objectif de rassembler les éléments les plus proches possibles selon un ou plusieurs critères. Il a également pour but de créer des ensembles les plus éloignés possibles, sur ce ou ces critères. Le critère utilisé dans notre cas sera l'homogénéité sémantique.

3.2 Regroupement des mots-clés par une adaptation de la méthode Hoffmann, Lomonosov et Sitharam (HLS)

Dans la mesure où elle est constituée d'un ensemble de phases paramétrables, la méthode d'Hoffmann, Lomonosov et Sitharam (HLS) telle que nous la présentons est très souple. Ce paramétrage nous permettra comme nous le verrons de supprimer ou conserver des liens en fonction de leur valeur relative au poids du mot.

3.2.1 Principes de base la méthode HLS

Cette méthode, proposée initialement par Hoffman C., Lomonosov N. et Sitharam M. (Hoffman *et al.*, 1997) est une méthode de décomposition structurelle ascendante. Elle recherche des ensembles d'objets rigides. Ces agrégats sont ensuite assemblés récursivement. La méthode est un exemple des méthodes de rigidification récursives de GCSP (Geometric Constraint Satisfaction Problem).

La méthode HLS comprend cinq phases. Une première phase d'analyse consiste à regrouper les agrégats. Cette phase d'analyse se compose de trois parties : fusion, extension et condensation. La phase de fusion recherche les agrégats de taille minimale. La phase d'extension consiste à inclure un objet voisin dans l'agrégat courant et ce, tant qu'il existe un objet voisin à insérer. L'opération d'extension utilise un opérateur d'extension qui va permettre d'étendre l'agrégat courant « A » aux objets *voisins*. La phase de condensation place les objets regroupés dans l'agrégat en cours de constitution et met à jour le plan d'assemblage. La phase d'Assemblage : la phase d'assemblage exécute un plan d'assemblage où chaque agrégat est considéré comme un objet de départ.

3.2.2 Mise en œuvre de la méthode de HLS

Dans notre étude, illustrée dans cette section, nous allons adapter la méthode HLS à notre contexte en définissant par exemple l'agrégat minimum comme une clique. La phase de fusion recherchera donc ces objets. L'opération d'extension utilise un opérateur d'extension suivant la règle suivante : « Le graphe de l'agrégat doit toujours rester bi-connexe pendant les opérations d'extension ».

Définition Graphe bi-connexe : Un graphe est bi-connexe si chaque point est relié par au moins deux chemins vers n'importe quel autre point du graphe. (cf. Figure 3).

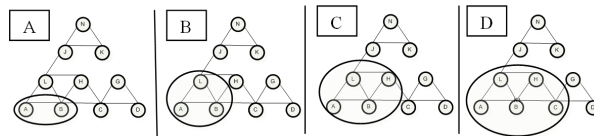


Figure 3. Illustration du déroulement de l'algorithme Fusion/Extension.

Nous nommerons *poids* le nombre de recherches liées à un objet. Cet objet étant soit un mot-clé, soit une relation *R* inter mots-clés. Le poids d'un mot-clé est donc le nombre de requêtes incluant ce mot-clé. Le poids d'une relation R_{AB} telle que $A R_{AB} B$ est le nombre de requêtes incluant les deux mots-clés A et B.

Le poids d'un mot-clé : Nb étant le nombre total de requêtes, MC_i étant l'élément de valeur vrai ou faux si le mot-clé est présent dans la requête (vrai valant 1, faux valant 0).

$$\text{Soit le poids d'un mot-clé A noté } P_A = \sum_{I=1}^{Nb} MC_I$$

Le poids d'une relation : Soit un mot-clé A et un mot-clé B, il existe une relation R_{AB} telle qu'il existe $A R_{AB} B$, Nb étant le nombre total de requêtes, R_I étant l'élément de valeur vrai ou faux si les mots-clés sont conjointement présents dans la requête (vrai valant 1, faux valant 0).

$$\text{Soit le poids d'une relation } R_{AB} \text{ noté } PR_{AB} = \sum_{I=1}^{Nb} R_I$$

Le poids total d'un mot-clé n'est pas nécessairement la somme des poids de ces relations. En effet, une même recherche peut inclure plusieurs mots-clés et donc compter pour « un » dans le poids du mot-clé (cf. figure 4).

Utilisation du poids pour l'orientation du graphe et amélioration de l'opérateur d'extension :

Nous proposons de compléter l'opérateur d'extension par une prise en compte de la notion de poids relatif. Il semble évident que le poids de la relation est à comparer aux poids des mots-clés en relation. Une relation d'un poids de « 1 » entre un mot-clé A pesant « 1000 » et un mot-clé B pesant « 2 » ne représente pas du tout la même importance relative. Ainsi la relation pèsera 10^{-3} du poids du mot-clé A et 0.5 du poids du mot-clé B. Afin de prendre en compte ce poids relatif nous orientons et pondérons le graphe de la matrice présenté en figure 4.. Nous utiliserons pour ceci la valeur du poids du mot-clé de départ sur le poids de la relation du mot-clé de départ avec le mot-clé cible. On notera ce rapport *CFL* (Coefficient de Fiabilité de Lien). Ainsi pour un mot-clé A en relation avec le mot-clé B noté $A R_{AB} B$, P_A le poids du mot-clé A, PR_{AB} le poids de la relation R_{AB} . Soit le *Coefficient de Fiabilité de Lien*, noté *CFL*, du mot-clé A vers le mot-clé B : $CFL_{A \rightarrow B} = P_A / PR_{AB}$

Matrice symétrique - graphe non dirigé						Matrice asymétrique - graphe dirigé - CRB (%)							
Mot	Poids	A	B	C	D	E	Mot	R_{CRB}	A	B	C	D	E
A	8	-	6	7	0	2	A	->	-	75	87.5	0	25
B	10	6	-	10	0	0	B	->	60	-	100	0	0
C	20	7	10	-	2	1	C	->	35	50	-	10	5
D	500	0	0	2	-	1	D	->	0	0	0.4	-	0.25
E	2	2	0	1	1	-	E	->	100	0	50	50	-

Figure 4. Matrice asymétrique d'un graphe orienté pondéré – CFL.

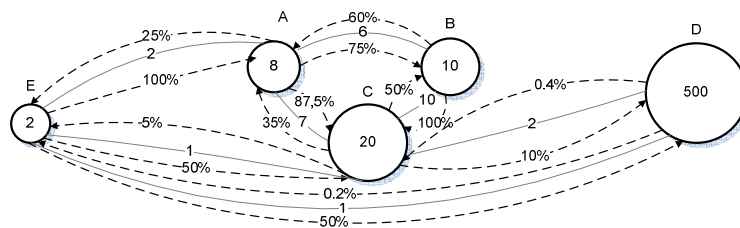


Figure 5. Graphe orienté pondéré du CFL de la matrice (donnée à titre d'exemple) présentée en figure 4 (CFL est ici présenté en % pour en faciliter la compréhension).

De façon à ne pas maintenir des liens présentant un *CFL* trop faible, nous ne prendrons en compte que les relations présentant un *CFL* supérieur à une valeur prédominée nommée Valeur Minimale de *CFL* ou *Val-Min-CFL* du poids du mot. De façon à ne pas perdre les mots de faible poids en relation avec des mots de poids fort nous maintiendrons quel que soit le *CFL* de sens inverse toutes relations ayant un *CFL* supérieur à la valeur d'activation ou *Val-Activ-CFL* du poids du mot

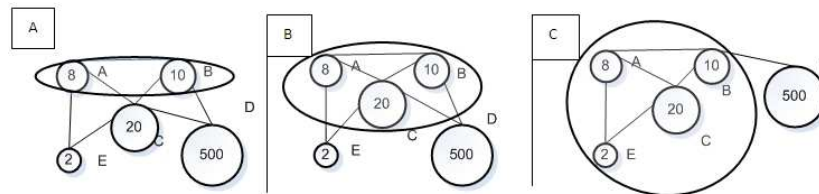


Figure 6. Illustration du déroulement de l'algorithme Fusion/Extension en utilisant l'Opérateur d'extension définitif.

L'opérateur d'extension définitif sera donc basé sur les règles suivantes : Le graphe doit rester bi-connexe, un *CFL* inférieur à *Val-Min-CFL* supprimera la relation sauf si le *CFL* de sens inverse est supérieur à *Val-Activ-CLF*

Dans l'exemple de la figure 5, la liaison C-D n'est pas maintenue car le $CFL_{D \rightarrow C}$ est inférieur au *Val-Min-CFL* fixé et le $CFL_{C \rightarrow D}$ est inférieur au *Val-Activ-CFL* fixé. L'élément « D » ne peut donc rejoindre l'agrégat car le graphe résultant ne serait alors plus bi-connexe.

Mécanisme de regroupement des mots-clés en agrégats :

Une paire de mots-clés constituant un diad ne peut appartenir au plus qu'à un agrégat. En effet, soit il existe un troisième mot-clé formant avec les deux premiers un triad et ce triad ne sera présent que dans un et un seul agrégat, soit il n'existe pas de triad incluant le diad et le diad n'est alors dans aucun agrégat (cf. Algorithme I.).

```

Pour chaque mot-clé X faire [Phase de fusion]
  Extraire les mots-clés Y qui forment un triad valide selon l'opérateur d'extension avec X
  Pour chaque couple de mots-clés X-Y valides faire [Phase d'extension]
    S'il n'existe pas d'agrégat contenant le couple X-Y et que le couple n'a pas été testé?
      Créer un nouvel agrégat « X-Y » et ajout de X-Y
      Tant que l'on ajoute des mots-clés dans l'agrégat faire
        Pour les mots de l'agrégat
          Rechercher de nouveaux mots en triad
          Ajouter des mots-clés formant le triad avec les mots de l'agrégat et
          Noter des couples trouvés comme « testés »
        Fin Pour
      Fin Tant que
    Fin Si
  Fin Pour [Fin de Phase d'extension]
Fin Pour [Fin de Phase de Fusion]
  
```

Algorithme 1. Regroupement des mots-clés en agrégats.

3.3 Postulat et technique de validation sémantique

3.3.1 Proposition d'une technique de validation sémantique

Les postulats à la proposition d'une technique de validation sémantique sont les suivants :

- Internet est majoritairement constitué de sites Web et de documents sémantiquement cohérents. Nous convenons qu'il existe des exceptions telles que des dictionnaires ou des listes d'objets en vente, mais les considérons comme numériquement faibles.

- les utilisateurs de moteurs de recherche sur Internet ont une conscience et une expérience suffisante pour utiliser des mots-clés ayant un lien entre eux et avec le sujet recherché.

Sur un ensemble de mots et de recherche suffisamment importants pour effectuer un traitement statistique, il devrait donc être possible d'observer un comportement différent, lorsque l'on compare le nombre de sites retournés par des requêtes utilisateurs à des requêtes combinant des mots de manière aléatoire.

Afin d'éclairer notre propos, nous soumettons en tant qu'utilisateur, trois recherches de trois mots-clés au moteur de recherche du site aol.com et une recherche combinant un mot-clé de chacune de ces recherches utilisateurs. Cette dernière étant notre recherche aléatoire.

Comme on peut le constater dans l'exemple Figure 6, trois mots-clés pris aléatoirement dans un ensemble de requêtes donnent des résultats significativement inférieurs en nombre de sites retournés à des requêtes plus « sémantiquement cohérentes » proposées par un utilisateur. Ceci n'a bien sûr de valeur que d'un point de vue statistique ; rien n'interdisant à un monsieur ou une madame « Besancenot » de placer une photo de sa personne sur Internet jouant du saxophone de la célèbre marque « Selmer » devant un plat d'épinard le tout accompagné d'une description.

ID	Requête	nb de sites retournés
1	+besancenot +état +france	164 000
2	+épinard +crème +beurre	37 400
3	+saxophone +selmer +jazz	66 300
4	+selmer + besancenot +épinard	0

Figure 6. Nombre de sites retournés par le moteur de recherche du site d'aol.com en fonction de la cohérence sémantique de celle-ci.

3.3.2 Technique de validation sémantique comparée

Notre but n'est pas de fournir une méthode de validation sémantique absolue, mais d'obtenir un indice de qualité sémantique. Cet indice est défini comme un ratio et n'a donc pas d'unité. Il n'est que le reflet de la comparaison comportementale des agrégats aux tests définis. Il permettra d'évaluer des méthodes de regroupement et leurs évolutions. Afin de créer cette mesure, nous proposons la comparaison du

nombre de sites, trouvés par le moteur de recherche du site aol.com, entre une référence, des combinaisons extraites des agrégats eux-mêmes et des combinaisons extraites d'un ensemble de mots-clés combinés au hasard. La taille de trois mots-clés représente la taille minimale d'un agrégat. Il est donc impossible de construire des recherches utilisant plus de trois mots-clés sans exclure de cette mesure les agrégats les plus petits. La validation des mots-clés par paire pourrait sans doute présenter un intérêt mais représenterait un nombre de combinaisons très important. Nous avons donc choisi de présenter les mots-clés trois par trois au moteur de recherche d'aol.com. Nous nommerons cet ensemble de trois mots-clés issus de l'agrégat : « triad ».

Toutes les combinaisons de trois mots-clés de chaque agrégat seront présentées au moteur de recherche. Cela représente 792756 combinaisons pour les agrégats sur l'échantillon d'étude. Chaque mot sera dans cette expérimentation précédé du signe "+" ce qui exclut les sites présentant les mots-clés inclus dans une chaîne de caractère de la liste des résultats. L'échantillon aléatoire a été formé de 500000 triads de mots-clés différents pris au hasard dans l'échantillon.

4. Expérimentation

4.1 Création de l'échantillon à étudier

Nous avons appuyé notre travail sur un extrait des fichiers de log du moteur de recherche aol.com. Ce fichier est mis à disposition du public par la société AOL à des fins d'étude. Il est disponible sur le site <http://gregsadetsky.com/aol-data>. L'extrait utilisé intègre trente trois millions de requêtes effectuées du 1 mars 2006 au 30 avril 2006. Ces requêtes sont principalement en langue anglaise. La structure du fichier intègre un identifiant, la date et l'heure de la recherche, le site éventuellement sélectionné ainsi que son rang (cf. figure 7).

AnonID	Query	QueryTime	temRank	ClickURL
142	rentdirect.com	2006-03-01 07:17:12		
142	www.prescriptionfortime.com	2006-03-12 12:31:06		
142	staple.com	2006-03-17 21:19:29		
142	staple.com	2006-03-17 21:19:45		
142	www.newyorklawyersite.com	2006-03-18 08:05:07		
142	westchester.gov	2006-03-20 03:03:09	1	http://www.westchestergov.com
142	space.comhttp	2006-03-24 20:51:24		

Figure 7. Fichier de log aol.com

Afin de travailler sur un échantillon représentatif et néanmoins manipulable, nous avons fait le choix de nous limiter à l'ensemble des requêtes d'une journée. La journée de référence choisie est aléatoirement celle du 17 avril 2006. Sur les requêtes de cette journée nous avons appliqué six règles : Les mots-clés sont définis comme un ensemble de lettres sans espace. Tout espace est donc lu comme un séparateur de mots-clés. Les guillemets ainsi que tous les éléments de ponctuation

ont été ignorés et remplacés par des espaces. Seuls les mots-clés possédant plus d'une lettre ont été conservés. Certains mots-clés jugés non significatifs ont aussi été écartés de l'étude. Seuls les mots-clés utilisés dans une requête ayant deux mots et plus ont été conservés.

Afin d'éviter de manipuler des mots au sens galvaudé par une trop grande utilisation, nous avons décidé de ne pas considérer les mots ayant été utilisés dans plus de 1000 recherches. Ecarter ces mots qui sont par définition les moins discriminants nous permet d'espérer éviter la construction de méga agrégats centrés sur ces mots-clés. Ces mots sont au nombre de 14 sur 51994 mots-clés étudiés soit 0.027 % de l'échantillon. Une fois ces différents « filtres » appliqués, l'objet de l'étude est un ensemble de 51980 mots-clés utilisés dans 200646 requêtes.

Après plusieurs essais sur des échantillons, nous avons défini comme valeurs possibles les seuils de Valeur Minimale de *CFL* ou *Val-Min-CFL* à 5 % du poids du mot-clé et la Valeur d'Activation ou *Val-Activ-CFL* à 20 % du poids du mot-clé. Ces valeurs pourront être modifiées lors de prochaines expérimentations, elles n'ont ici que valeur d'exemple et ne sont pas le sujet de l'étude.

4.2 Résultats et analyse de la technique de validation sémantique comparée

La démarche implantée a permis de former 9556 agrégats construits avec 38621 mots-clés dont 24537 mots-clés différents dans l'ensemble des agrégats. Le nombre moyen de mots-clés par agrégat est de 4,04. L'agrégat le plus important est de 133 mots-clés (cf. figure 8).

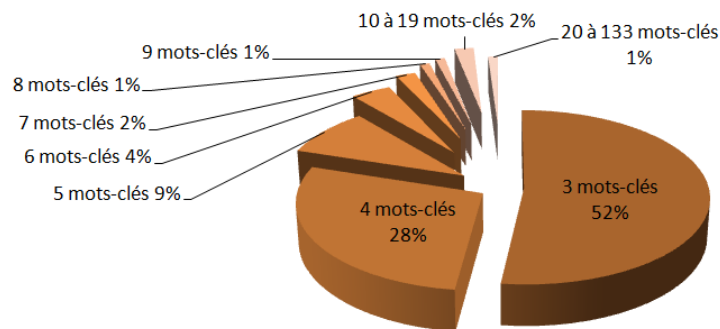


Figure 8. Répartition des agrégats en fonction du nombre de mots-clés.

4.2.1 Recherche d'un élément de comparaison

Une représentation graphique du nombre de sites retournés en fonction d'une population se heurte à des difficultés. L'étendue des valeurs de retour et le nombre de valeurs différentes retournées sont trop considérables pour en proposer une vision graphique. Dans notre cas nous allons de « 0 » site retourné à plus de 99 millions de sites pour certaines requêtes.

Pour pallier à ces difficultés nous représenterons les résultats dans une échelle semi-logarithmique en utilisant un regroupement des valeurs dans des classes. Un repère semi-logarithmique est un repère dans lequel l'un des axes, ici celui des ordonnées (y), est gradué selon une échelle linéaire alors que l'autre axe, ici celui des abscisses (x), est gradué selon une échelle logarithmique. L'avantage d'une représentation semi-logarithmique est son aptitude à représenter des mesures qui s'étalent sur des valeurs extrêmement larges. Des représentations semi-logarithmiques en puissance de 2 ont déjà été utilisées par Zipf (Zipf , 1935) dans ces études sur l'occurrence des mots à l'intérieur d'un texte.

Axe des ordonnées : Nous placerons en ordonnées le pourcentage de combinaisons trouvées par classe par rapport à l'ensemble des classes.

Axes des abscisses : Pour pouvoir comparer les résultats obtenus nous avons regroupé le nombre de sites retournés dans des classes exprimées dans un espace logarithmique. Afin de rester sur des classes les plus fines possibles nous avons choisi des classes par puissance de 2.

Nous comparons ici les deux courbes de réponses des deux espaces les plus éloignés sémantiquement selon le postulat posé au paragraphe 3.3.1. Nous comparerons la courbe issue des mots combinés aléatoirement avec la courbe de référence issue du test de triads pour laquelle il existe au moins une recherche incluant ces trois mots-clés. A l'exception des triads aléatoires, les autres triads testés sont extraits d'agrégats obtenus par la méthode HLS-CFL. La comparaison s'effectue ici dans une première phase de manière graphique (cf. figure 9). Nous remarquons 4 zones clairement identifiables (cf. figure 9), la zone A de 0, La zone B de 2^1 à 2^9 , la zone C de 2^{10} à 2^{20} et la zone D supérieure à 2^{20}

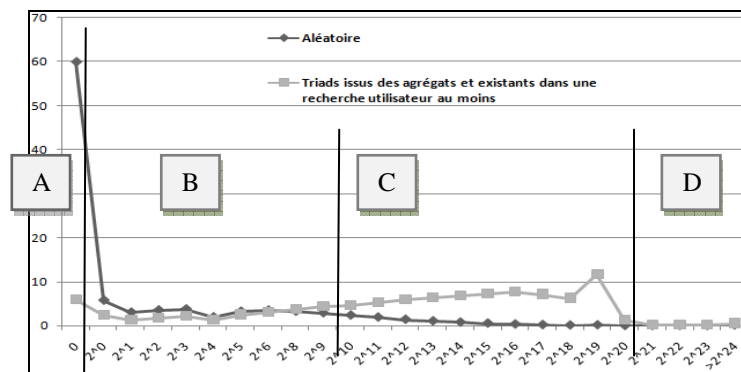


Figure 9. Comparaison des deux courbes les plus éloignées sémantiquement et détermination des zones remarquables.

Les zones « B » et « D » ne représentent pas beaucoup d'intérêt, les courbes ne présentant pas de différence notable. La zone « A » est limitée à une seule valeur et

ne peut donc représenter une étendue suffisante pour mener notre étude. La zone « C » est la zone la plus différente et d'une plage suffisante pour avoir un sens. Afin de mieux percevoir l'importance de la « C », reprenons une lecture du graphique en omettant les zones A B et D .

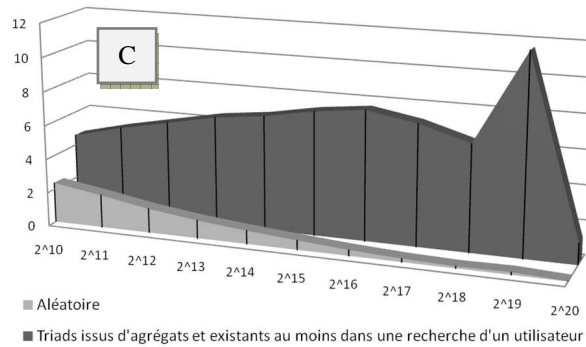


Figure 10. Zoom sur la zone « C » sélectionnée comme zone d'étude.

La zone « C » nous servira de zone de validation sémantique. Afin de pouvoir élaborer une comparaison rapide et arithmétique nous allons définir un coefficient.

4.2.2 Calcul du CVSC (Coefficient de Validation Sémantique Comparée)

Nous considérerons que les classes en puissance de deux forment une échelle d'indice « un » et comparons l'aire prise par les deux histogrammes. Le CVSC, ou Coefficient de Validation Sémantique Comparée, ayant alors la valeur « 1 » pour l'équivalence de l'histogramme des triads (de trois mots-clés) ayant été au moins une fois utilisés dans une même recherche et 0 pour la valeur de l'histogramme des triads aléatoires.

La formule mathématique sera donc pour une courbe particulière X à valider. On aura alors, pour une courbe X, :

$$CVSC_X = (A_X - A_A) / (A_R - A_A)$$

Où A_R définit l'aire de l'histogramme des triads dont tous les mots sont inclus au moins une fois tous ensembles dans une recherche telle que: $A_R = \sum_{i=10}^{20} Y_i = 70,24$

Où A_A définit la valeur de l'aire de l'histogramme des triads aléatoires telle que: $A_A = \sum_{i=10}^{20} Y'_i = 8,95$

Où A_X définit la valeur de l'aire de l'histogramme des triads à comparer telle que: $A_X = \sum_{i=10}^{20} Y''_i$

4.2.3 Comparaison des coefficients CVSC pour des agrégats de tailles différentes

Afin de déterminer une cible pour des travaux futurs, il semble important de borner vers le haut la taille des agrégats. Nous allons comparer ici les CVSC pour des agrégats de taille différente. Nous remarquons rapidement - que ce soit de manière graphique (cf. figure 11) ou par le calcul du CVSC - que plus le nombre de mots-clés est important plus le CVSC a tendance à baisser.

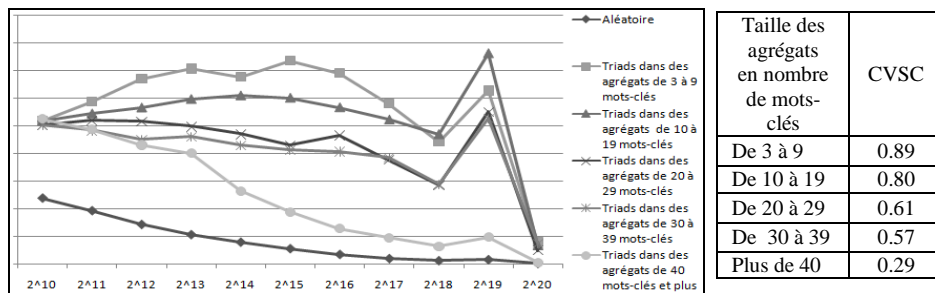


Figure 11. Comparaison du CVSC en fonction de la taille des agrégats en Zone « C » de validation sémantique.

Cette courte étude nous permet de constater que les agrégats d'une taille supérieure à 30 mots possèdent un CVSC inférieur ou égal à 0.5. Il semble donc que la taille de 30 à 40 mots soit statistiquement une cible à considérer comme un maximum pour garder une certaine cohérence sémantique.

5. Conclusion

Afin d'assurer de nouveaux services de type réseaux sociaux aux utilisateurs des moteurs de recherche, nous avons proposé un système de regroupement et de validation sémantique de mots-clés issus des fichiers de logs des moteurs de recherche. Pour illustrer cette approche nous avons développé l'exemple d'un nouveau service de construction de communautés dynamiques.

Afin de valider cette approche d'agrégation de mots-clés nous proposons une technique comparative qui repose sur une mesure sémantique de ces agrégats en exploitant entre autres les outils de recherche sur Internet comme validateurs. Cette technique peut servir à affiner et comparer les algorithmes de regroupement.

La méthode de regroupement que nous avons retenue et adaptée HLS-CFL peut encore évoluer et être améliorée. La démarche de validation sémantique comparée permettra alors d'arbitrer la qualité de ces évolutions.

Dans cette première approche, nous avons considéré les mots-clés comme des objets neutres et indépendants. Dans le futur, en utilisant l'algorithme de Porter et des outils intégrant Wordnet ou des dictionnaires ontologiques, des dictionnaires de synonymes ou des communautés issus de travaux tels que ceux de B. Gome (Gome, 2004), il sera possible de faire évoluer les algorithmes vers plus d'efficacité en recombinaison des agrégats de petites tailles. Il est aussi possible de repérer au sein des agrégats des ensembles « extrêmement liés » qui peuvent servir de noyaux à des méthodes complémentaires de réduction des agrégats de taille supérieure à 30 mots-clés. La réduction des agrégats de taille importante et supérieure à 30 mots-clés pourra se faire aussi par une modification de l'indice *CFL* (Coefficient de Fiabilité de Lien) et une mise en quarantaine mieux contrôlée des mots-clés sur-utilisés ou vides. Enfin une lecture des caractéristiques (coefficient de classification, distance moyenne, diamètre) des graphes créés par les agrégats et leur confrontation à la technique de validation peut aussi représenter une piste d'amélioration.

Le regroupement de mots-clés utilisés par les utilisateurs de moteurs de recherche sur Internet a plusieurs objectifs. Une fois validés comme sémantiquement cohérents, ils peuvent servir par exemple à déterminer le profil d'un utilisateur. Ainsi un utilisateur proposant des mots-clés pourra, si ses propres mots-clés sont liés à un agrégat, se voir proposer de nouveaux services : proposition de sites repères, amélioration de la procédure de recherche par la proposition de mots complémentaires et une mise en contact immédiate avec des utilisateurs ayant les mêmes centres d'intérêt. Autant de nouvelles pistes offertes par ces agrégats obtenus par l'application de notre approche.

6. Références

- Balfe, E., 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.
- Berge B., 1958, *Théorie des graphes et ses Applications*, Dunod.
- Bruno, G., Fabien, M., 2008. *From Random Graph to Small Word by Wandering*, eprint arXiv:0804.0149.
- Cui, H., Wen, J., Nie, J., et Ma, W., 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., Golgher, P.B., de Moura, E.S., Ziviani, N., 2003. *Using association rules to discover search engines related queries*. First Latin American Web Congress (LAWEB'03), pp. 66–71.
- Fu, L., Goh, D.H-L., Foo, S. S-B., et Na, J-C., 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., Goh, D.H-L., Foo, S. S-B., et Supangat, . Y., 2004. *Collaborative querying for enhanced information retrieval*, 2004 European conference on research and advanced technology for digital libraries, pp. 378–388.
- Hoffman, C, Lomonosov, A. et Sitharam, M, 1997. *Finding Solvable Subsets of Constraint Graphs. International Conference on Principles and Practice of Constraint Programming*, LNCS 1330, Berlin, Springer-Verlag, pp. 463–477.
- Hoffman, C, Jaon-Arinyo, 1997. *Symbolic constraints in constructive geometric constraint solving. Journal of Symbolic Computation*, 23:287–299.
- Hoffman, C, Lomonosov, A. et Sitharam, M, 2000, *Planning Geometric constraint decomposition via optimal graph transformations*. Actes de la conférence AGTIVE'99, LNCS 1779, Springer-Verlag, pp. 309–324
- Hoffman, C, Lomonosov, A. et Sitharam, M, 2001. *Decomposition plans for geometric constraint systems*, Part II: New Algorithms, Academic Press. *J. Symbolic Computation* (2001): 31, pp. 409–427.
- Koutsoupias, 2000, N.: *Exploring web access logs with correspondence analysis*. Methods Bransford, J.D., Brown, A.L.O., & Cocking, R.R. (Eds.).
- Jermann, J., 2002, *Résolution de contraintes géométriques par rigidification récursive et propagation d'intervalles*, Thèse de Doctorat, Université de Nice Sophia-Antipolis, pp. 104, pp.121–160
- Jermann, C., Neveu, B., et Trombettoni, G, 2004. *Algorithmes pour la détection de rigidités dans les CSP géométriques. Issue spéciale du Journal Électronique d'Intelligence Artificielle (JEDAI-JNPC'03)*.
- Latapy, M., 2007. *Grands graphes de terrain – mesure et métrologie, analyse, modélisation, algorithmique*. H.D.R., Université Pierre et Marie Curie, Paris, France.
- Luce, R.D., Perry, A.D. , 1949, *A Method of matrix analysis of group structure*, *Psychometrika*. 14, pp. 95–116
- Newcomb, T. M., Turner, R. H., Converse, P. E., 1965, *Psychology: The Study of Human Interaction*. New York: Holt, Rinehart & Winston.
- Ohkubo, M., Sugizaki, M., Inoue, T., Tanaka, K.,, 1998. *Extracting information demand by analyzing a www search log*. *Information Processing Society of Japan Journal* 39(7), pp. 2250–2258.
- Reinhard, D., 2005, *Graph Theory Electronic*, Springer-Verlag Heidelberg.
- Shingo, O., Masaru, K., 2006. *Clustering of Search Engine Keywords Using Access Logs*, Conference on Database and Expert Systems Applications, DEXA 2006, LNCS 4080, Springer-Verlag Berlin Heidelberg, pp. 842–852.
- Thibaut, J. W., Kelley, H. H., 1959, *The Social Psychology of Groups*. New York: Wiley.
- Zipf, G. K., 1935, *The Psychobiology of Language, an Introduction to Dynamic Philology*, Boston, Houghton-Mifflin.