Agrégats de mots-clés validés sémantiquement

Pour de nouveaux services d'accès à l'information sur internet

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

christian@belbeze.com, {Max.Chevalier, Chantal.Soule-Dupuy}@irit.fr

RÉSUMÉ. *A l'heure du web social, nous présentons une solution destinée à définir de nouveaux services tels que la construction automatique et dynamique de communautés d'utilisateurs : l'agrégation de mots-clés. Ces agrégats de mots-clés sont issus des recherches antérieures des utilisateurs réalisées au travers d'un moteur de recherche. 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. A des fins d'évaluation et de validation, nous proposons de comparer les résultats obtenus par le moteur de recherche à partir des agrégats de mots-clés ainsi créés et de définir un coefficient de cohérence sémantique de ces agrégats. Nous mesurons dans une expérimentation la perte de cohérence sémantique liée à l'augmentation de la taille des agrégats. L'intérêt de notre approche réside dans le fait qu'elle peut être considérée comme une brique de base pour un grand nombre de systèmes « communautaires » et ainsi exploitée pour offrir encore plus de services à l'usager.*

ABSTRACT. *At the hour of the social Web, we present a solution being able to be used as a basis for the definition of new services such as automatic and dynamic construction of communities of users: the aggregation of keywords. These aggregates of keywords result from former research of the users of a search engine. We present the process which we followed to obtain an algorithm for gathering keywords coming from log files (tracing files); we illustrate this algorithm through its application to a log file of the search engine aol.com. At ends of evaluation and validation, we propose to compare the results obtained by the search engine starting from the aggregates of keywords thus created and to define a semantic coefficient of coherence of these aggregates. We measure in an experimentation the loss of semantic coherence related to the increase in size of the aggregates. The interest of the approach suggested lies in the fact that it can be regarded as a basis for a great number of "community" systems and thus exploited to offer even more services to the user.*

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

KEYWORDS: *keywords, aggregate, cluster, semantic, graph, log files, search engine, group, bag of words, internet.*

DOI:10.3166/DN.12.1.81-105

1. Introduction

Lors de travaux précédents au travers notamment du projet SiSSI (Belbèze et Soulé-Dupuy, 2007), nous avons observé les difficultés rencontrées par les internautes recherchant des informations. Au-delà des difficultés de manipulation des outils de recherche et du repérage de l'information pertinente au sein des pages, l'évaluation de la confiance à accorder à un site ou à un document est un souci présent. Le manque de compétence propre à un domaine de recherche ou à son contexte linguistique est une des raisons d'échec. Ainsi, la prise de conscience qu'il n'est plus seul devant sa page web et ainsi 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 autant de 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. Une des solutions possibles est donc la construction automatique et dynamique de communautés d'usagers, c'est-à-dire que les usagers seront reliés à différentes communautés représentant leurs centres d'intérêt, qui dans notre cas sont caractérisés par des agrégats de mots-clés. Ces relations entre usagers et communautés pourront évoluer dans le temps.

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. Le terme généralement consacré au regroupement de mots-clés est celui de « cluster » (grappe). La notion de « cluster » fait à la fois référence aux nœuds d'un réseau et à la structure porteuse de ce réseau. Dans notre cas, le processus de regroupement, bien qu'utilisant les liaisons comme ressources, génère une simple liste. C'est pourquoi nous avons préféré le terme d'agrégat. Un agrégat est défini par (Bayaly et Cunny, 1986) comme un ensemble de nœuds liés logiquement dans un graphe.

Afin d'identifier les communautés d'usagers basées sur les centres d'intérêt, nous proposons donc de regrouper des mots-clés issus de recherches d'information dans des agrégats présentant une forte cohérence sémantique. Nous entendons par cohérence sémantique la capacité d'un groupe de mots à recouvrir un champ d'un domaine le plus précis possible. Cette liste de mots pourrait s'apparenter à ce qu'en lexicologie on nomme un champ lexical. Mel'Cuk *et al.* (1995) donnent du champ lexical la définition suivante : « Nous appelons champ lexical d'un champ sémantique l'ensemble des vocables dont les lexies de base appartiennent à ce champ sémantique ».

A l'usage, à partir des mots-clés utilisés lors d'une recherche d'information par un nouvel usager, nous pourrons (1) identifier les communautés pertinentes grâce aux agrégats ainsi construits afin de (2) rapprocher cet utilisateur des utilisateurs attachés aux communautés les plus proches et ainsi (3) lui offrir de nouveaux services basés sur ces communautés.

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 (1997), appelée « méthode HLS ». 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 repose sur un espace de mots-clés issus d'extraits de logs du moteur de recherche d'aol.com. Les agrégats construits sont basés sur la cooccurrence de mots-clés dans les différentes requêtes issus du fichier log d'un moteur de recherche. Enfin, la validation réalisée repose sur la comparaison des réponses du moteur de recherche auquel nous soumettons des agrégats de mots issus d'ensembles de mots-clés créés aléatoirement et les agrégats de mots construits selon l'approche proposée. Cette validation expérimentale nous permettra de mieux définir les limites de notre approche et d'y apporter de futures évolutions.

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 au travers notamment des adaptations de la méthode HLS que nous avons proposées ainsi que la technique de validation sémantique associée. La section 4 illustre l'expérimentation que nous avons menée à partir d'extraits du fichier de logs d'aol.com ainsi que les résultats positifs obtenus. Enfin, la dernière section (section 5) dresse un bilan ainsi qu'un panorama des perspectives d'évolution de l'approche proposée.

2. Contexte et état de l'art

La « clusterisation » 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 requêtage 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 issus de moteurs de recherche sur internet. Par la suite, nous nous focaliserons sur les travaux relatifs à la création d'agrégats sémantiquement homogènes qui ont inspiré nos travaux.

Certains travaux réalisés sur le regroupement 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. Ainsi, O. Shingo *et al.* (2006) proposent la création de clusters construits par l'association de mots-clés ayant amené des utilisateurs à consulter le même site. Ces ensembles de quelques mots sont alors additionnés pour former des clusters plus larges si les sites sélectionnés par les internautes sont dans la même communauté de sites (une communauté de sites est un ensemble de sites reliés par des liens http). Cette solution présente l'avantage d'une double base de construction, les utilisateurs et les documents servant de sources. Mais elle ne semble pas possible sur l'ensemble du web sans une mise en œuvre de moyens gigantesques qui permettrait la construction de l'ensemble des communautés des sites de tout internet. De plus, l'apparition de nouveaux clusters de mots-clés ou des clusters de mots-clés de faible usage est ralentie par le fait que les mots-clés se doivent d'être trouvés dans des documents déjà indexés. Notre méthode va bien utiliser les moteurs de recherche dans une phase de validation mais uniquement d'un point de vue statistique, ce qui ne gênera pas la création d'agrégats utilisant de nouvelles associations de mots-clés.

D'autres travaux, comme ceux de (Cui *et al.*, 2002) et de (Fonseca *et al.*, 2004) tentent de créer des clusters de mots-clés en corrélant les mots-clés utilisés dans la recherche avec ceux mis en avant par les URL retournés (URL, titre, mots-clés cités dans la page html…) et sélectionnés par l'internaute. Cette méthode présente les mêmes dépendances aux index des moteurs de recherche que la précédente. De plus, cette méthode pose aussi le problème de la pertinence de l'ordonnancement des résultats par les moteurs de recherches commerciaux d'internet, l'ordre de présentation influant fortement la sélection d'un site proposé par l'utilisateur.

Les travaux de (Koutsoupias, 2000), quant à eux, ont eu pour but de créer une technique d'enrichissement de la requête en proposant à l'internaute un complément de mots-clés issus de requêtes fréquentes ou de requêtes retournant un grand nombre de sites. Ces services sont aujourd'hui offerts en standard par certains moteurs de recherche (Yahoo, Google). Ces techniques très pratiques ne sont cependant utilisées que pour la création de groupes de quelques mots, généralement trois à cinq.

La création d'agrégats de mots-clés sémantiquement homogènes semble plus appropriée à la proposition de nouveaux services aux internautes, notamment pour la définition de communautés d'usages, que les systèmes de clusterisation précédemment présentés. Dans notre représentation graphique les éléments de liaison correspondront à l'« utilisation conjointe » de mots, les mots étant eux-mêmes les nœuds du graphe. La mise en œuvre de graphes représentant des mots-clés est illustrée par exemple par le site http://adlab.microsoft.com/Vnext/Entity-Association-Graph/Default.aspx (cf. figure 1).

Il existe plusieurs méthodes de regroupements d'objets dans des graphes. Seules les techniques permettant un recouvrement des différents agrégats sont à considérer. De par sa nature, un mot doit être rattachable à des contextes multiples.

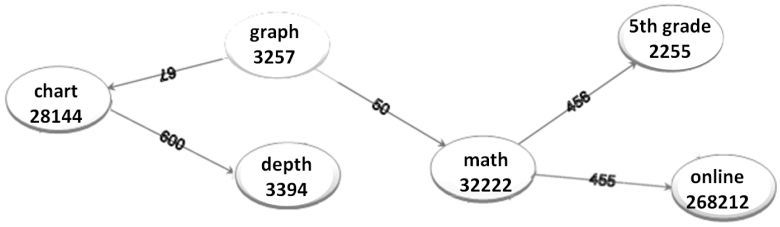

Figure 1. *Exemple d'un graphe « Entity Association Graph » visualisable sur le site http://adlab.microsoft.com/Vnext/Entity-Association-Graph/*

Une première catégorie de méthodes s'intéresse aux structures appelées « cliques ». La définition de la *clique* a été originellement posée par (Festinger, 1949) ainsi que (Luce et Perry, 1949). Telle que définie par (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. Cependant, au sein de notre échantillon, l'obtention d'agrégats uniquement basés sur cette figure n'est pas apparue pertinente, les cliques ne dépassant pas alors un niveau de 9 éléments. La méthode basée sur l'agrégation de cliques développée par Palla (Palla *et al.*, 2006) pour contourner ce problème est sans doute une piste à considérer. Palla propose de combiner les cliques de telle sorte qu'une clique de n sommets soit rattachée à une autre dans un agrégat si n-1 sommets sont communs. Si cette technique a donné de bons résultats dans l'univers des réseaux sociaux et de la biologie moléculaire, elle a aussi été appliquée à la création d'agrégats de mots (Chavalarias *et al.*, 2008). Mais son relatif manque de souplesse nous a fait préférer une méthode plus ouverte et plus paramétrable.

Une deuxième catégorie de méthodes de regroupement, toujours à base de graphes, repose sur des approches de résolution de contraintes géométriques. Un système de contraintes géométriques se compose d'un ensemble d'objets géométriques soumis à des contraintes géométriques. Résoudre un système de contrainte géométrique consiste à fournir une position, une orientation et des dimensions à chacun de ses objets géométriques de sorte que toutes les contraintes géométriques soient satisfaites. (Jian-Xin Ge1 *et al.* 1999) définissent la résolution d'un système de contraintes en utilisant des représentations graphiques comme la transformation du système de contraintes en un graphe puis par la recherche de séquences de construction issues de l'analyse du graphe. Les types d'objets et de contraintes géométriques dépendent du domaine d'application considéré (mécanique, dessin, biologie, littéraire…). Un ensemble d'objets est défini comme rigide s'il est indéformable, autrement dit si les objets n'acceptent plus de déplacement entre eux. Ici les systèmes rigides représentent un ensemble de nœuds ou un agrégat répondant aux contraintes du système. Les travaux

de (Hoffman *et al.*, 1997) et de Jermann *et al.* (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 cette section, 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 reste un champ d'investigation encore largement ouvert.

3. Agrégats de mots-clés : vers la proposition de nouveaux services aux utilisateurs de moteur de recherche sur internet

L'observation d'internautes recherchant des informations à travers un moteur de recherche (Belbèze et Soulé-Dupuy, 2007) nous a permis de déterminer que plus l'utilisateur possédait une connaissance approfondie du sujet traité, plus il lui était facile de trouver l'information manquante recherchée. Ainsi un utilisateur expérimenté peut taper les mots d'une chanson pour en trouver le texte complet. Cette compétence de base déterminante pour l'accès à l'information est un élément rapidement transmissible *via* une mise en relation efficace des usagers, d'où notre proposition d'agrégation de mots-clés issus des moteurs de recherche afin en particulier de construire des communautés dynamiques d'usagers.

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 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'autodescription. Les efforts d'authentification sont d'autant plus mal acceptés qu'ils ne correspondent, le plus souvent, qu'à un espace réservé (pour un usage personnel uniquement). Le temps passé à une autodescription, lui aussi, ne correspond pas aux habitudes d'immédiateté des services les plus utilisés tels que les moteurs de recherche.

La création de communautés dynamiques pourrait donc être exploitée afin de permettre 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 n'en 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 l'exemple présenté dans les figures 2 et 3, Georges recherche, de façon anonyme, des sites web en utilisant les mots-clés B, E, G et H. Grâce à l'agrégat N°1 contenant ces mots, Georges peut se voir proposer de rentrer en contact avec des utilisateurs ayant des centres d'intérêts proches des siens. Il peut soit ouvrir un salon de discussion où seront invités automatiquement les internautes concernés par les mots-clés de l'agrégat N°1, soit démarrer une conversation en messagerie instantanée avec l'utilisateur « Anonyme » ou bien laisser un message à Annie (cf. figure 3).

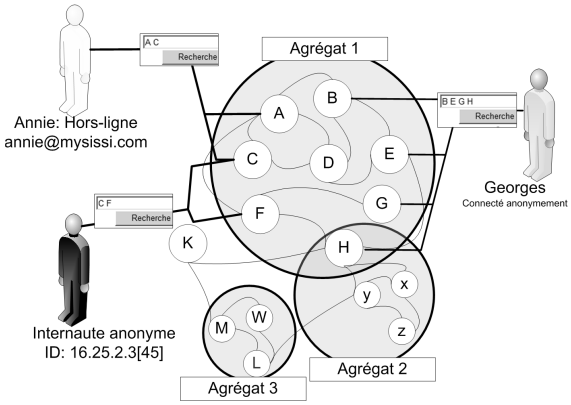

Figure 2. *Attachement des internautes à un agrégat en fonction de leurs recherches*

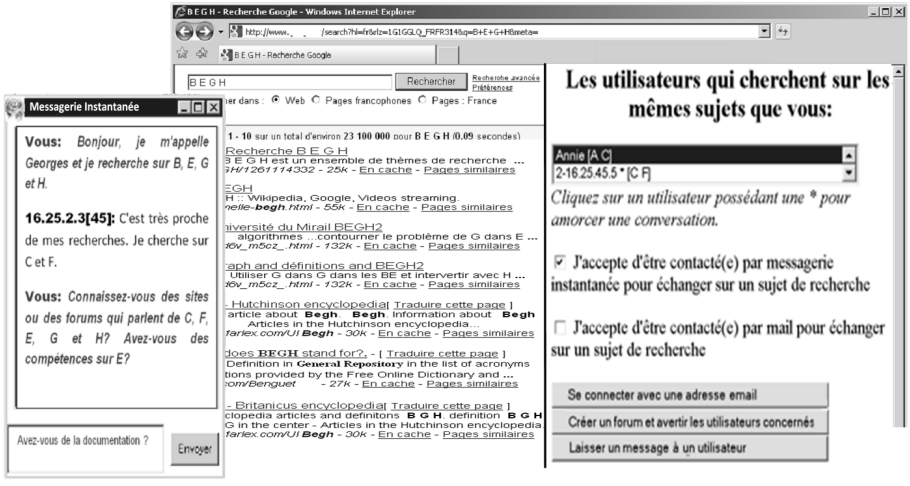

Figure 3. *Utilisation des nouveaux services de communication au sein de la communauté dynamique*

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 prédominant utilisé dans notre cas sera l'homogénéité sémantique.

3.2. *Agrégation des mots-clés : 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) est très souple. Ce paramétrage nous permet de supprimer ou conserver des liens entre des mots-clés en fonction de leurs valeurs elles-mêmes relatives aux poids des mots.

3.2.1. *Principes de base de la méthode HLS*

Cette méthode, proposée initialement par (Hoffman *et al.*, 1997) est un exemple de méthode de rigidification récursive de GCSP (*Geometric Constraint Satisfaction Problem*). Plus exactement une méthode de décomposition structurelle ascendante. Elle recherche des ensembles d'objets rigides. Ces agrégats sont ensuite assemblés récursivement.

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 repose sur un opérateur d'extension). La phase de condensation place les objets regroupés dans l'agrégat en cours de constitution et met à jour un plan d'assemblage. Enfin la phase d'assemblage exécute le plan d'assemblage où chaque agrégat est considéré comme un objet de départ.

3.2.2 *Mise en œuvre de la méthode HLS*

Pour nos besoins, nous adaptons la méthode HLS à notre contexte en définissant par exemple l'agrégat minimum comme une clique. La phase de fusion recherche donc ces objets. L'opération d'extension utilise un opérateur d'extension que nous définissons selon la règle suivante : « Le graphe de l'agrégat doit toujours rester bi-connexe pendant les opérations d'extension » afin de préserver un niveau de rigidification de l'agrégat. On dit qu'un graphe est bi-connexe si chaque point est relié par au moins deux chemins à n'importe quel autre point du graphe.

La figure 4 illustre les phases de fusion et d'extension que nous avons adaptées dans la suite de cette section.

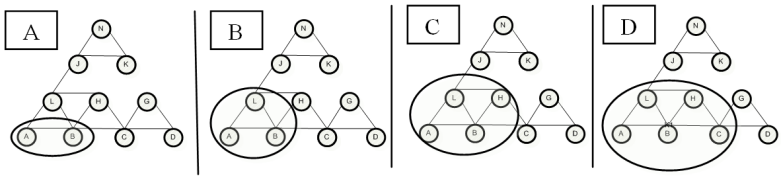

Figure 4. *Illustration du déroulement de l'algorithme fusion/extension*

Ces phases reposent sur la notion de *poids*. Dans notre cas, le poids correspond au nombre de recherches liées à un objet. Cet objet étant soit un mot-clé issu des requêtes, soit une relation *R* inter mots-clés (cf. ci-après). Le poids d'un mot-clé correspond au nombre de requêtes utilisant ce mot-clé. Le poids d'une relation R_{AB} telle que A R_{AB} B correspond au nombre de requêtes incluant les deux mots-clés A et B.

Le poids d'un mot-clé : soit *Nb* le nombre total de requêtes, MC_I l'élément de valeur vrai si le mot-clé est présent dans la requête (vrai valant 1), ou faux sinon (faux valant 0), alors le poids d'un mot-clé A noté P_A est calculé comme suit :

$$P_A = \sum_{I=1}^{Nb} MC_I$$

Le poids d'une relation : soit deux mots-clés A et B, une relation R_{AB} telle que A R_{AB} B, *Nb* le nombre total de requêtes, R_I l'élément de valeur vrai si les mots-clés sont conjointement présents dans la requête (vrai valant 1), ou faux sinon (faux valant 0), alors le poids d'une relation R_{AB} noté PR_{AB} est calculé de la façon suivante :

$$PR_{AB} = \sum_{I=1}^{Nb} R_I$$

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 5).

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 tableau 1. Nous utilisons 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), le *coefficient de fiabilité de lien*, noté *CFL*, du mot-clé A vers le mot-clé B noté $CFL_{A=>B}$ est calculé comme suit : $CFL_{A=>B} = P_A / PR_{AB}$ *avec* P_A poids du mot-clé A, PR_{AB} le poids de la relation R_{AB}.

Matrice symétrique - graphe non dirigé								**Matrice asymétrique - graphe dirigé - *CFL* (%)**						
Mot	**Poids**	**A**	**B**	**C**	**D**	**E**		**Mot**	***Relation***	**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	-

Tableau 1. *Matrice asymétrique d'un graphe orienté pondéré – CFL*

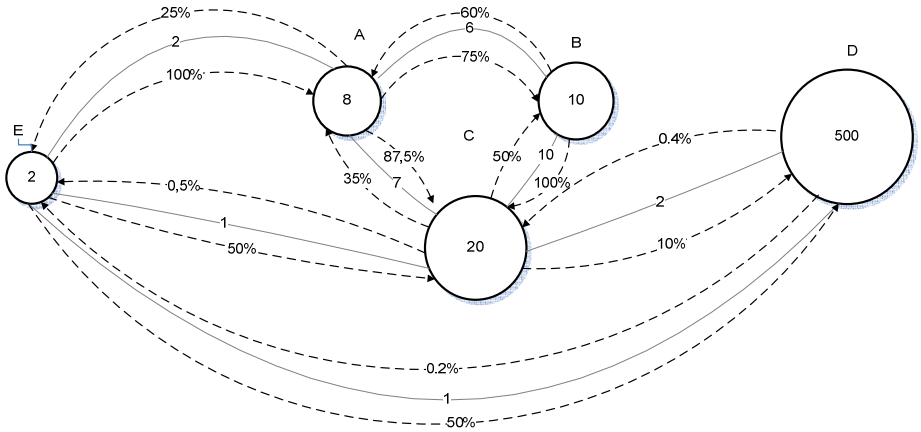

Figure 5. *Exemple de graphe orienté pondéré du CFL correspondant aux données de la matrice présentée en tableau 1.(CFL est ici présenté en % pour en faciliter la compréhension)*

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 prédéfinie 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 toutes relations ayant un *CFL* supérieur à la valeur d'activation ou *Val-Activ-CFL* du poids du mot et ce quel que soit le CFL de sens inverse.

Des valeurs seuils pour *Val-Min-CFL et Val-Activ-CFL* ont été fixées par expérimentations respectivement à 5 % et 20 % (cf. section 4.1).

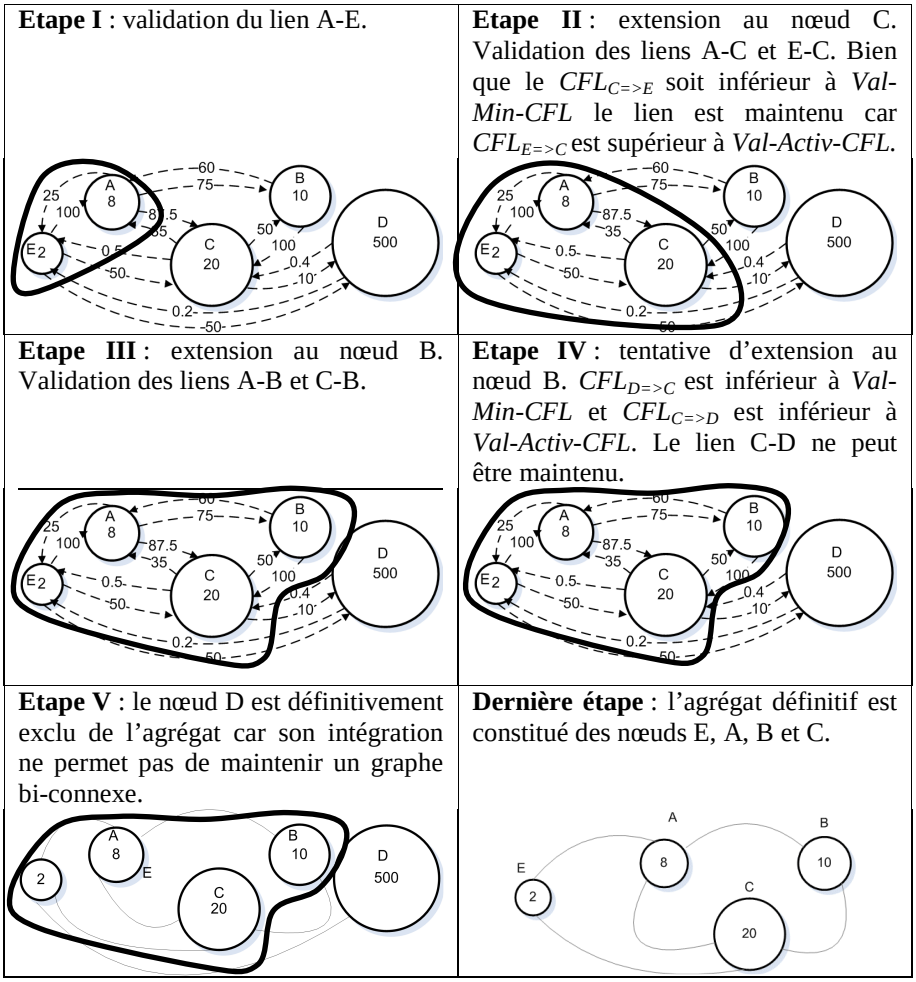

Figure 6. *Illustration du déroulement de l'algorithme fusion/extension en utilisant l'opérateur d'extension proposé*

L'opérateur d'extension définitif que nous proposons est 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 6, la liaison C-D n'est donc pas maintenue car le $CFL_{D=>C}$ est inférieur au *Val-Min-CFL* fixé et le $CFL_{C=>D}$ 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 donc une diade, ne peut appartenir au plus qu'à un agrégat. En effet, soit il existe un troisième mot-clé formant avec les deux premiers mots-clés une triade et cette triade ne sera présente que dans un et un seul agrégat, soit il n'existe pas de triade incluant la diade et la diade n'est alors dans aucun agrégat (cf. Algorithme 1).

Pour chaque mot-clé X faire *[Phase de fusion]*
  Extraire les mots-clés Y formant une triade 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 triade
          Ajouter des mots-clés formant la triade avec les mots de l'agrégat
          Noter des couples trouvés comme « testés »
        Fin Pour
      FinTant que
    Fin Si
  Fin Pour *[Fin de Phase d'extension]*
Fin Pour *[Fin de Phase de Fusion]*

Algorithme 1. *Algorithme général de regroupement des mots-clés en agrégats*

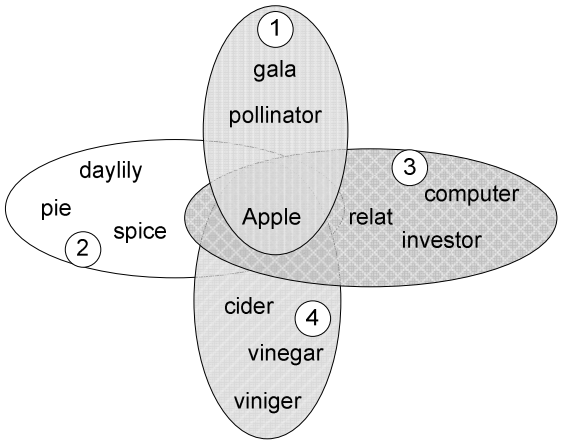

Figure 7. *Exemple de 4 agrégats partageant le même mot commun « Apple » résultant de notre proposition*

A titre d'exemple et afin d'éclairer le lecteur sur les résultats que la technique d'agrégation permet d'obtenir, nous proposons ici une représentation schématique des différents agrégats générés incluant le mot « apple ».

Comme on peut le remarquer dans la figure 7, les 4 agrégats sont cohérents et illustrent quatre contextes (acceptions) différents identifiés par rapport au mot-clé « Apple » (pomme). Ainsi, l'agrégat 1 fait référence au fruit (pomme) lui-même, le 2 à la botanique avec le lys (« daylily ») ayant pour nom « Apple Pie Spice », le 3 à la marque d'ordinateur bien connue, et enfin le 4 au cidre de pomme.

Afin de valider les résultats de notre approche, nous proposons d'appliquer une mesure de validation sémantique sur les agrégats ainsi obtenus.

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 des ensembles de mots et de recherches 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 les exemples illustrés dans le tableau 2., 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'épinards.

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

Tableau 2. *Nombre de sites retournés par le moteur de recherche du site d'aol.com en fonction de la cohérence sémantique de la requête*

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 de comparer le nombre de sites retrouvés par le moteur de recherche du site aol.com à partir de requêtes basées sur des combinaisons extraites des agrégats eux-mêmes avec le nombre de sites retrouvés (par le même moteur de recherche) à partir de requêtes basées sur des combinaisons aléatoires de mots-clés (combinaisons indépendantes des agrégats construits).

Trois mots-clés représentent la taille minimale d'un agrégat (triade). 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 trop important. Nous avons donc choisi de présenter les mots-clés au moteur de recherche d'aol.com par triade.

Toutes les combinaisons de trois mots-clés de chaque agrégat seront présentées au moteur de recherche. Cela représente 792 756 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 caractères de la liste des résultats. L'échantillon aléatoire a été formé de 500 000 triades de mots-clés différents pris au hasard dans l'échantillon.

4. Expérimentation

4.1. *Création de l'échantillon à étudier*

Nous avons effectué nos expérimentations à partir d'un extrait des fichiers de log du moteur de recherche aol.com. Ce fichier est mis à disposition du public 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 8).

AnonID	Query	QueryTime	itemRank	URL
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 8. *Extrait du 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 de recherches. La journée de référence choisie aléatoirement est celle du 17 avril 2006.

Sur les requêtes de cette journée, nous avons appliqué les six règles suivantes :

– 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 (cf. tableau 3),

– 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 filtré les mots-clés ayant été utilisés dans plus de 1 000 recherches (cf. tableau 4.). Ecarter ces mots, qui sont par définition les moins discriminants, nous permet d'éviter la construction de méga agrégats centrés sur ces mots-clés. Ces mots sont au nombre de 14 sur 51 994 mots-clés étudiés soit 0,027 % de l'échantillon. Une fois ce filtre appliqué, l'ensemble de mots-clés exploité dans notre expérimentation contient 51 980 mots-clés utilisés dans 200 646 requêtes.

.com	be	don't	having	http	l.	off	this	where	your
al	been	el	he	if	la	she	to	who	yourself
all	by	elle	her	Il	like	so	too	why	
alt	can	for	here	in	on	st	us	will	
and	com	from	his	my	our	st.	was	with	
are	de	had	is	ne	ours	than	we	www	
as	do	how	it keep	no	out	th	what	www.	
at	does	href		of	re	their	when	you	

Tableau 3. *Liste des mots-clés exclus de l'étude en tant que mots-clés non significatifs*

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 constituent pas le sujet de l'étude.

Mots-clés	Poids	Mots-clés	Poids	Mots-clés	Poids	Mots-clés	Poids
sale	1011	city	1273	sex	1560	new	2413
york	1071	tax	1458	lirycs	1561	free	3956
bank	1083	State	1532	country	1884		
home	1139	school	1539	pictures	2020		

Tableau 4. *Mots-clés exclus car utilisés dans plus de 1 000 requêtes le 17/04/06*

4.2 *Résultats et analyse de la technique de validation sémantique comparée*

La démarche implantée a permis de former 9 556 agrégats construits avec 38 621 mots-clés dont 24 537 mots-clés différents dans l'ensemble des agrégats (cf. figure 9). Le nombre moyen de mots-clés par agrégat est de 4,04. L'agrégat le plus important est de 133 mots-clés.

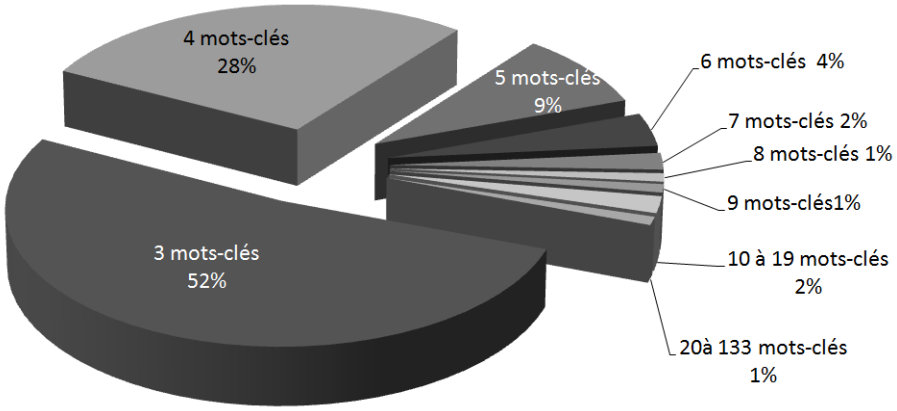

Figure 9. *Répartition des agrégats en fonction du nombre de mots-clés*

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 quelques 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 ces difficultés, nous représenterons les résultats selon 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 (1935) dans ces études sur l'occurrence des mots à l'intérieur d'un texte.

Ainsi, dans la figure 10, l'axe des abscisses est gradué en puissances de 2. En effet, pour pouvoir comparer les résultats obtenus, nous avons regroupé le nombre de sites retournés dans des classes exprimées dans un espace logarithmique. Si les échelles logarithmiques sont habituellement en puissance de 10, afin de présenter une échelle plus détaillée, nous avons choisi des classes par puissance de 2. L'axe des ordonnées représente alors le pourcentage de combinaisons trouvées par classe par rapport à l'ensemble des classes.

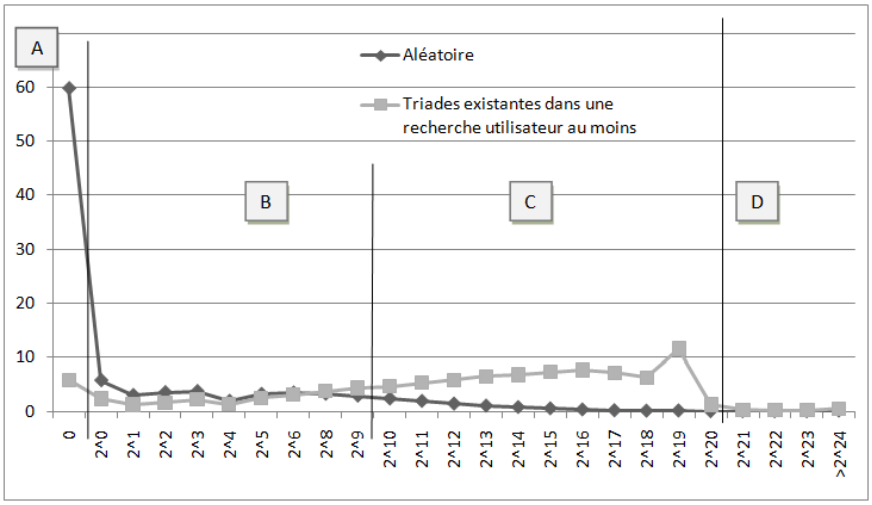

Figure 10. *Comparaison des deux courbes les plus éloignées sémantiquement et détermination des zones remarquables*

Nous comparons ici les deux courbes de réponses des deux espaces les plus éloignés sémantiquement selon le postulat posé en section 3.3.1. Nous comparerons la courbe issue des mots combinés aléatoirement (excluant des triades de mots utilisés dans une recherche) avec la courbe de référence issue du test de triades pour laquelle il existe au moins une recherche incluant ces trois mots-clés. A l'exception des triades aléatoires, les autres triades testées sont extraites d'agrégats obtenus par la méthode HLS-CFL. La

comparaison s'effectue ici, dans une première phase, de manière graphique. Sur la figure 10, nous distinguons 4 zones clairement identifiables, la zone A de 0, la zone B de 2^1 à 2^9, la zone C de 2^10 à 2^20 (cf. figure 11) et la zone D supérieure à 2^20.

Les zones « B » et « D » ne présentent 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.

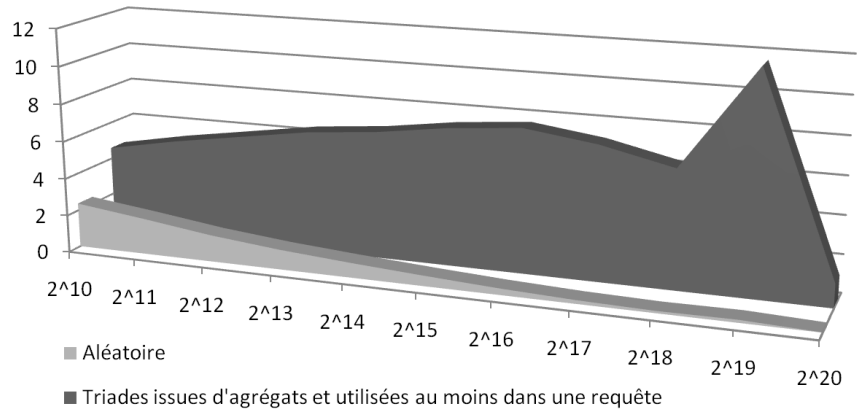

Figure 11. *Zoom sur la zone « C » sélectionnée comme zone d'étude*

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

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 triades (de trois mots-clés) ayant été au moins une fois utilisées dans une même recherche et 0 pour la valeur de l'histogramme des triades aléatoires.

La formule mathématique sera donc définie 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 triades dont tous les mots sont inclus au moins une fois tous ensembles dans une recherche :

$$A_R = \sum_{i=10}^{20} Yi = 70{,}24$$

où A_A définit la valeur de l'aire de l'histogramme des triades aléatoires :

$$A_A = \sum_{i=10}^{20} Y'i = 8{,}95$$

où *Ax* définit la valeur de l'aire de l'histogramme des triades à comparer :

$$A_X = \sum_{i=10}^{20} Y''i$$

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 tailles différentes. Nous remarquons rapidement - que ce soit de manière graphique (cf. figure 12) ou par le calcul du *CVSC* - que plus le nombre de mots-clés est important, plus le CVSC a tendance à baisser.

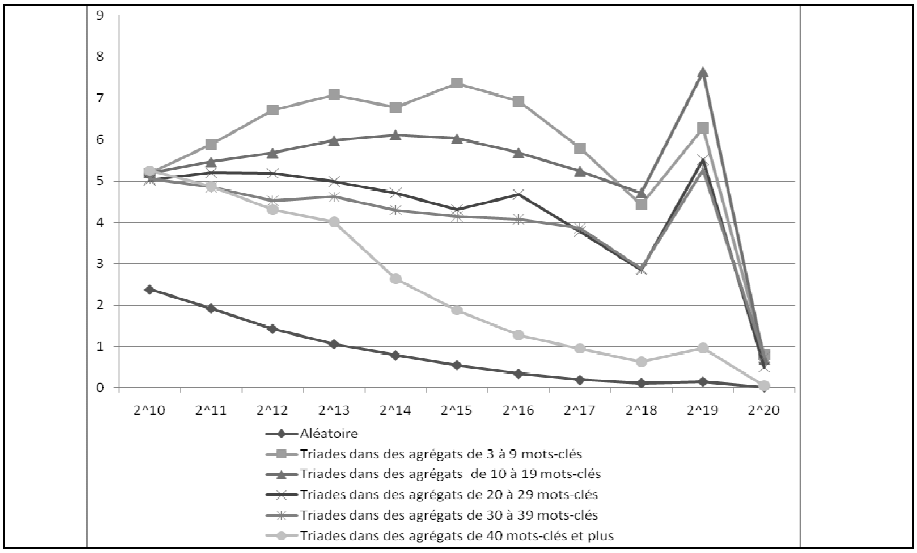

Figure 12. *Représentation graphique des CVSC en fonction de la taille des agrégats en zone « C » de validation sémantique*

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.

Taille des agrégats en nombre de mots-clés	CVSC
De 3 à 9	**0.89**
De 10 à 19	**0.80**
De 20 à 29	**0.61**
De 30 à 39	**0.57**
Plus de 40	**0.29**

Tableau 5. *Valeurs des CVSC en fonction de la taille des agrégats en zone « C » de validation sémantique*

4.2.4. *Comparaison des coefficients CVSC en excluant les recherches utilisateurs*

Afin d'estimer la perte de cohérence sémantique liée à la notion d'agrégats, il nous a semblé pertinent de comparer les coefficients sémantiques obtenus pour les mêmes classes d'agrégats en excluant les triades utilisées dans une recherche au moins.

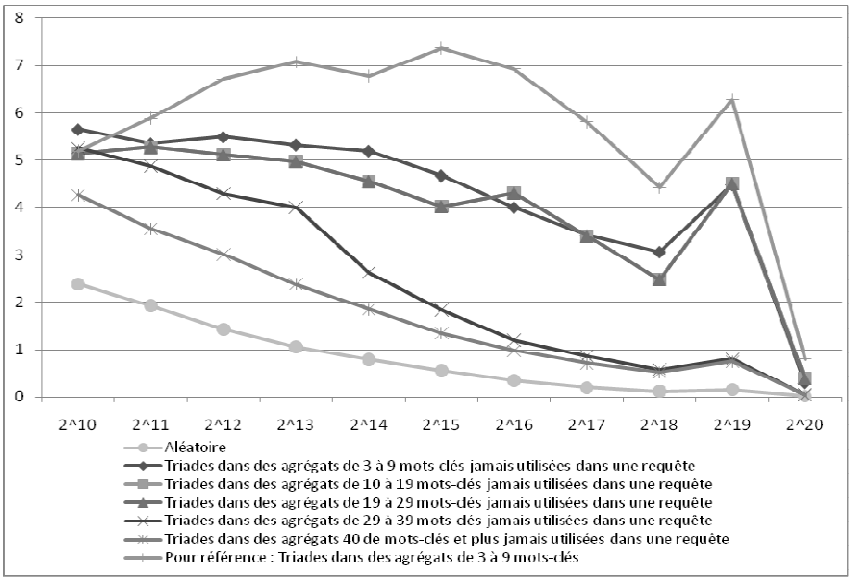

Figure 13. *Représentation graphique des CVSC en fonction de la taille des agrégats en zone « C » de validation sémantique en excluant les triades incluses dans une requête d'utilisateur*

L'observation des chiffres du CVSC des triades issues d'agrégats et n'ayant jamais été utilisées dans une recherche par un utilisateur nous conforte fortement sur le chiffre à ne pas dépasser. En effet, les agrégats de moins de 30 mots gardent un ratio supérieur à la moyenne.

Il est difficile de déterminer sans une étude détaillée au cas par cas les raisons de la baisse du coefficient. Cependant, la possibilité qu'un mot soit utilisé dans des acceptions différentes peut en être une des causes.

Taille des agrégats en nombre de mots-clés	CVSC	Perte
De 3 à 9	**0.62**	**0.27**
De 10 à 19	**0.57**	**0.23**
De 20 à 29	**0.56**	**0.05**
De 30 à 39	**0.28**	**0.29**
De 40 à 49	**0.17**	**0.12**

Tableau 6. *Valeurs des CVSC en fonction de la taille des agrégats en Zone « C » de validation sémantique en excluant les triades utilisées dans une requête*

4.3. *Exemples d'agrégats intégrant un mot commun*

Dans cette section, nous illustrons, au travers de deux exemples, deux agrégats centrés sur un mot ayant plusieurs acceptions afin d'illustrer la baisse du coefficient identifiée précédemment.

L'agrégat de la figure 14 ci-après illustre les concepts de musique et de cuisine, notamment au travers du mot « chef ». Ainsi différentes acceptations de ce mot interviennent dans cet agrégat. Cependant, lors de l'évaluation de la cohérence sémantique de cet agrégat, le tirage aléatoire des mots-clés dans l'agrégat risque de générer un certain nombre de triades ayant une faible cohérence sémantique. En voici trois exemples :

1) *+nettoyage +musique +orchestre*
2) *+cuisine +chef + musique*
3) *+piano +nettoyage +sol*

A titre d'exemple supplémentaire, prenons un autre agrégat de mots : *abiline, arunde, arundl, aubun, avalanche, b2600, car, cars, chevrolet, dealerships, electronic, fj40, fordsale, gaffn, hamptonroad, ignition, lexus, lynchb, maine, microwave, murrieta, outboard, parts, pax, selecti, ulster, uplander, used, usedfront, virgini, waterville*.

Cet agrégat a été construit grâce notamment à :

– la requête utilisateur *+used +car +pax* qui renvoie 284 000 sites : Pax est une référence de pneu de marque Michelin et d'autres pièces détachées ;

– la requête utilisateur *+used +car +abiline* qui renvoie 1 140 sites : *Abiline* est un centre de vente et d'achat de pièces détachées ;

– la requête utilisateur *+used +car +murietta* qui renvoie 17 100 sites : *Murieta* est un centre de réparation de véhicules.

Ces trois requêtes utilisateurs sont toutes situées dans la zone C. Cependant, la requête aléatoire issue de cet agrégat, utilisée dans la mesure de la cohérence sémantique, *+abiline +murietta +Pax* (où *Abiline* devient un prénom, *Murietta* le nom d'une ville et Pax le mot latin signifiant Paix) ne retourne qu'un seul site qui se trouve être présent dans la zone C !

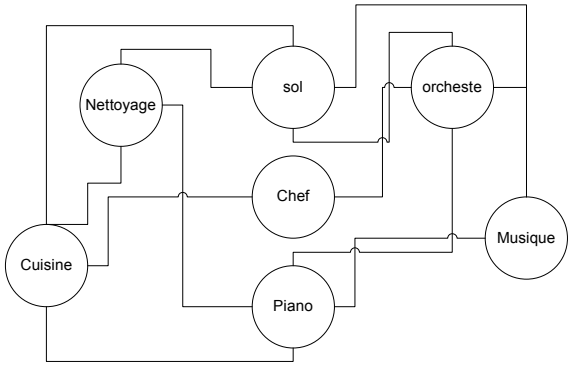

Figure 14. *Exemple d'agrégat intégrant des mots ayant plusieurs acceptions (musique/cuisine)*

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 Kleinberg *et al.* (2000) ou ceux plus récents de M. Latapy (2007), de B. Gaume *et al.* (2008), il sera possible de faire évoluer les algorithmes vers plus d'efficacité en recombinant 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 également se faire 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. Ces agrégats peuvent également servir à déterminer de nouveaux usages linguistiques ainsi qu'à déterminer des contextes de traduction. Ils peuvent de plus être utilisés comme éléments déterminant dans la veille technologique ou commerciale pour surveiller l'apparition ou l'évolution de nouveaux centres d'intérêt. Autant de nouvelles pistes offertes par ces agrégats obtenus par l'application de notre approche.

6. Bibliographie

Bailey D. A., Cuny J. E., *Graph grammar based specification of interconnection structure for massively parallel computation*, Lecture Notes in Computer Science 291 Grapg Grammars and their Application to Computer Science 3rd International Workshop, Warrenton, Virginia, USA, 1986, p. 73-85.

Balfe E., Smyth B., *A Comparative Analysis of Query Similarity Metrics for Community-Based Web Search*. *ICCBR 2005*, H. Munoz-Avila and F. Ricci (Eds.), LNAI 3620, 2005, p. 63-77.

Belbèze C., Soulé-Dupuy C., « Apport des services Web dans l'amélioration de l'accès à l'information sur le Web ? », *Actes de la Conférence en Recherche d'Information et Applications*, Conférence francophone en Recherche d'Information et Applications (CORIA 2007), Saint-Etienne, Association Francophone de Recherche d'Information et Applications (ARIA), 2007, p. 35-51.

Berge B., *Théorie des graphes et ses Applications*, Dunod, 1958.

Chavalarias D., Cointetj J.-P., *Bottom-up scientific field detection for dynamical and hierarchical science mapping - methodology and case study*, Scientometrics, vol. 75, n° 1, 2008, (DOI): 10.1007/s11192-007-1825-6.

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*, p. 325-332.

Festinger L., The analysis of sociograms using matrix algebra. *Human Relations*, vol. 2, n°2, 1949, p. 153-158.

Fonseca B.M., Golgher P.B., de Moura E.S., Ziviani N., "Using association rules to discover search engines related queries", *First Latin American Web Congress (LAWEB'03)*, 2003, p. 66-71.

Fu L., Goh D.H-L., Foo S. S-B., et Na J.-C., *Collaborative querying through a hybrid query clustering approach*, Digital libraries, Technology and management of indigenous knowledge for global access, ICADL, 2003, p. 111-122.

Fu L., Goh D.H.-L., Foo S. S-B., et Supangat Y., "Collaborative querying for enhanced information retrieval", *European conference on research and advanced technology for digital libraries*, 2004, p. 378-388.

Gangnet M. and Rosenberg B., 1993, "Constraint programming and graph algorithms", *Annals of Mathematics and Artificial Intelligence*, vol. 8, n° 3-4, p. 271-284.

Gaume B., Mathieu F., *From Random Graph to Small Word by Wandering*, eprint arXiv:0804.0149, april 2008.

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, p. 463-477.

Hoffman C., Jaon-Arinyo, R, 1997. "Symbolic constraints in constructive geometric constraint solving", *Journal of Symbolic Computation*, 23, p. 287-299.

Hoffman C., Lomonosov A. et Sitharam M., « Planning Geometric constraint decomposition via optimal graph transformations", *Actes de la conférence AGTIVE'99*, LNCS 1779, Springer-Verlag, 2000, p. 309-324.

Hoffman C., Lomonosov A. et Sitharam M., *Decomposition plans for geometric constraint systems*, Part II: New Algorithms, Academic Press. J. Symbolic Computation, 31, 2001, p. 409-427.

Jermann J., Résolution de contraintes géométriques par rigidification récursive et propagation d'intervalles, Thèse de Doctorat, Université de Nice Sophia-Antipolis, 2002, p. 104 ; 121-160.

Jermann C., Neveu B., et Trombettoni G., « Algorithmes pour la détection de rigidités dans les CSP géométriques », *Journal Électronique d'Intelligence Artificielle* (JEDAI-JNPC'03), 2004.

Kleinberg J., « The Small-World Phenomenon: An Algorithmic Perspective", *Proceedings of the 32nd ACM Symposium on Theory of Computing*, 2000.

Koutsoupias N., *Exploring web access logs with correspondence analysis*, Methods Bransford, J.D. Brown, A.LO., & Cocking, R.R. (eds.) 2000.

Latapy M., Grands graphes de terrain – mesure et métrologie, analyse, modélisation, algorithmique. H.D.R., Université Pierre et Marie Curie, Paris, France, 2007.

Luce R.D., Perry A.D., *A Method of matrix analysis of group structure*, Psychometrika, 14, 1949, p. 95-116.

Mel'Cuk J., Clas A., et Polguère A., *Introduction à la lexicologie explicative et combinatoire*, Edition Duculot, 1995, p. 176-179.

Newcomb T. M., Turner R. H., Converse P. E., *Psychology: The Study of Human Interaction*, New York, Holt, Rinehart & Winston, 1965.

Ohkubo M., Sugizaki M., Inoue T., Tanaka K., "Extracting information demand by analyzing a www search log", *Information Processing Society of Japan Journal*, vol. 39, n° 7, 1998, p. 2250-2258.

Pallal G., Imre D., Tamás V., *The Critical Point of k -Clique Percolation in the Erdős–Rényi Graph*, Springer Netherlands, vol. 128, 2007, p. 219-227.

Reinhard D., *Graph Theory Electronic*, Springer-Verlag Heidelberg, 2005.

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, p. 842-852.

Thibaut J. W., Kelley H. H., *The Social Psychology of Groups*, New York, Wiley, 1959.

Zipf G. K., *The Psychobiology of Language, an Introduction to Dynamic Philology*, Boston, Houghton-Mifflin, 1935.