île Rousse 2005

Journée sur les systèmes d'information élaborée

UNE ANALYSE DES REQUETES D'UN MOTEUR INTRANET - VERS UNE AMELIORATION DU SYSTEME D'INFORMATION

Bertrand Delecroix (*) – Renaud Eppstein (*)
bertrand.delecroix@wanadoo.fr , eppstein@univ-mlv.fr

(*) ISIS/CESD, Université de Marne la Vallée, Cité Descartes, 5 Boulevard Descartes, Champs-sur-Marne, 77454 Marne-la-Vallée Cedex 2

Mots clefs :
Fouille de données textuelles, recherche d'information, Analyse de log, intelligence économique

Résumé
Cet article présente les premiers résultats d'une étude menée sur l'ensemble des requêtes soumises au moteur de recherche du serveur de veille concurrentielle, financière et commerciale de France Télécom.

Depuis plusieurs années, de nombreuses expériences ont tenté de mettre en avant l'inadéquation qui existe entre les moteurs de recherche et le comportement, les besoins et la connaissance des utilisateurs. Il est aisé de comprendre que plus les utilisateurs ont des profils larges et variés, et plus les sujets couverts par une collection de données sont divers, plus il sera difficile d'adapter les outils de recherche de façon à ce qu'ils fournissent les documents les plus appropriés.

Dans cet article, et sur la base de l'analyse des requêtes soumises à un moteur de recherche indexant une base documentaire spécialisée, nous démontrerons que même dans le contexte d'un intranet, et avec une base documentaire qualifiée et destinée à des "utilisateurs avancés", l'inadéquation précédemment citée existe toujours.

Cet article va présenter l'analyse des logs de requêtes de l'Arianet, serveur intranet d'information commerciale, financière et marketing. Le fichier des logs est constitué des 220 000 requêtes soumises au moteur de recherche durant l'année 2003. La base documentaire interrogée est constituée d'environ 600 000 documents, traitant, au sens large, du domaine des télécommunications et des nouvelles technologies.

Après une rapide présentation des logs analysés, nous analysons de façon plus spécifique l'utilisation des mots clé et la structure des requêtes. Nous proposons un premier ensemble de résultats qui améliorent notre compréhension du comportement et des besoins des utilisateurs. Nous proposons également une interprétation de certaines statistiques, spécialement celles concernant le très petit nombre de requêtes ayant généré au moins un clic et ayant amené à la visualisation d'un document. Ces observations nous amèneront enfin à proposer un ensemble de recommandations pour améliorer le système d'information.

1. Introduction

L'analyse du comportement de l'utilisateur revêt une importance particulière. En effet, c'est en connaissant parfaitement comment l'utilisateur va élaborer ses stratégies de recherche d'information, et en prenant ainsi conscience de ses réussites et de ses échecs, qu'il sera possible de lui proposer des outils susceptibles d'améliorer significativement ses recherches, et donc son accès à l'information.
L'utilisateur de l'ARIA a accès à un fonds documentaire très volumineux, constitué de 600 000 documents. De plus, chaque jour, trois à quatre mille dépêches alimentent le flux quotidien de nouveaux documents disponibles.
Ces documents sont accessibles de plusieurs façons. Depuis la page d'accueil (Figure 1) de l'Arianet, les documents sont tout d'abord accessibles via une classification thématique, effectuée de façon automatique (A). Une navigation par source est également mise à disposition des utilisateurs (C). La page d'accueil du site propose également une sélection de documents, mis en avant par les experts de l'Aria. Enfin, il est possible d'effectuer des requêtes via le moteur de recherche (A). Le moteur de recherche, Search'97 de Verity, indexe l'intégralité de la base documentaire disponible à l'Aria. Cependant, les utilisateurs sont nombreux à se plaindre de la piètre qualité des résultats que leur fournit le moteur de recherche. Nous avons donc voulu, à partir de l'historique des requêtes dont nous disposions, analyser le comportement des utilisateurs.

Figure 1. : Page d'accueil de l'Arianet

Pour cela, nous avons analysé l'ensemble des requêtes soumises au moteur de recherche. Ces analyses statistiques nous ont tout d'abord révélé les limites du moteur de recherche. D'autre part, nous avons pu voir que les utilisateurs maîtrisent très imparfaitement la syntaxe reconnue par le moteur de recherche utilisé.
Nous allons donc dans un premier temps présenter la structure des données analysées, avant de présenter les résultats concernant l'analyse du comportement de l'utilisateur face au moteur de recherche, puis tirer les enseignements et les conclusions que l'on peut tirer de cette analyse.

2. Description des données

L'ensemble des requêtes

L'analyse porte sur les requêtes soumises au moteur de recherche de janvier 2003 à novembre 2003. Durant ces onze mois, 218 693 requêtes en tout ont été soumises au moteur de recherche. Sur ces 218 693 requêtes, on distingue 46 210 requêtes différentes. Ainsi, en moyenne, une même requête est soumise 4,7 fois au moteur de recherche sur la période étudiée. Cependant, la dispersion est énorme : la requête la plus fréquemment soumise au moteur est *flarion* (3 024 fois) et on dénombre 25 911 requêtes qui ont été soumises une seule fois au moteur de recherche sur l'année.

Les requêtes identifiées

On appelle *requêtes identifiées* les requêtes qui, suite aux résultats ramenés par le moteur de recherche, ont suscité au moins un clic par leur utilisateur, et qui aura dû pour cela s'identifier. Au total, 26 383 requêtes identifiées sont soumises au moteur de recherche dont 16 840 requêtes différentes. Ainsi, en moyenne, une requête est soumise 1,5 fois au moteur de recherche. La requête identifiée la plus fréquente est *wifi*, soumise au moteur de recherche 1 266 fois, et cliquée 187 fois.

Les requêtes jamais identifiées

29 370 requêtes différentes ne vont jamais susciter de clics. Ces requêtes représentent un total de 81 366 requêtes. La requête jamais identifiée la plus fréquente est *network soma*, soumise 1 004 fois au moteur de recherche.
Pour résumer, si l'on considère les requêtes différentes, on a en tout 46 210 requêtes différentes, dont 16 840 vont susciter au moins un clic, et 29 370 ne vont jamais susciter de clics.

3. Etude du comportement utilisateur face au moteur de recherche

L'intensité des visites sur le moteur au cours des mois

Les chiffres généraux concernant les utilisateurs de l'Aria sont les suivants :

- Environ 8 000 identifiants et mots de passe ont été distribués ;
- En moyenne, 2 000 utilisateurs différents visitent et s'identifient chaque mois sur le site de l'Arianet.
- Ce chiffre, estimé sur l'année, s'élève à 4 000.

Ces chiffres concernent les utilisateurs accédant à des documents sur le site de l'Arianet, quels que soient les moyens d'accès à l'information.
Or, si l'on s'intéresse uniquement aux utilisateurs utilisant le moteur de recherche, ces chiffres sont très différents. En effet, sur l'année, 2 166 utilisateurs vont accéder à des documents *via* le moteur de recherche.

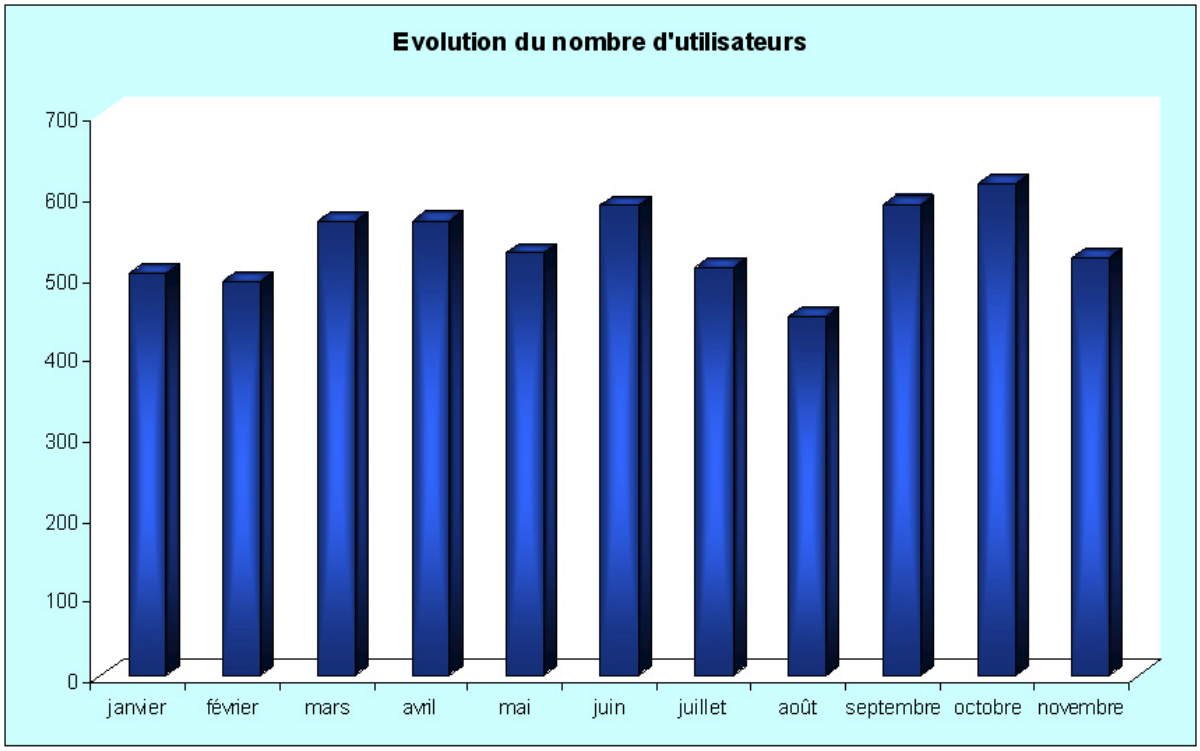

Figure 2. : Evolution du nombre d'utilisateurs du moteur de recherche

Le nombre d'utilisateurs identifiés sur le moteur de recherche varie de 449 (en août) à 615 (en octobre). Le nombre d'utilisateurs moyens par mois est de 538 ; l'écart-type est faible, à 47,7. Ainsi, si chaque mois, 2 000 utilisateurs accèdent à des documents sur le serveur de l'Aria, ils sont seulement 538 en moyenne, soit environ un quart, à le faire *via* le moteur de recherche. Les autres accèdent donc aux documents en utilisant d'autres moyens mis à disposition (navigation thématique, navigation par source…).

Le nombre de termes par requête

Sur l'ensemble de l'année, le nombre de termes utilisés pour constituer une requête varie de un à dix-huit. Quel que soit le type de requêtes (ensemble des requêtes, requêtes identifiées, ou requêtes jamais identifiées), les requêtes les plus fréquentes sont toujours les requêtes à deux termes, dont la fréquence est très légèrement supérieure à celle des requêtes à un terme. A eux deux, ces deux types de requêtes constituent presque les trois quarts des requêtes. Les requêtes à trois termes constituent de quinze à vingt pour cent des requêtes. Les requêtes constituées de quatre termes ou plus sont marginales, et rapidement décroissantes.

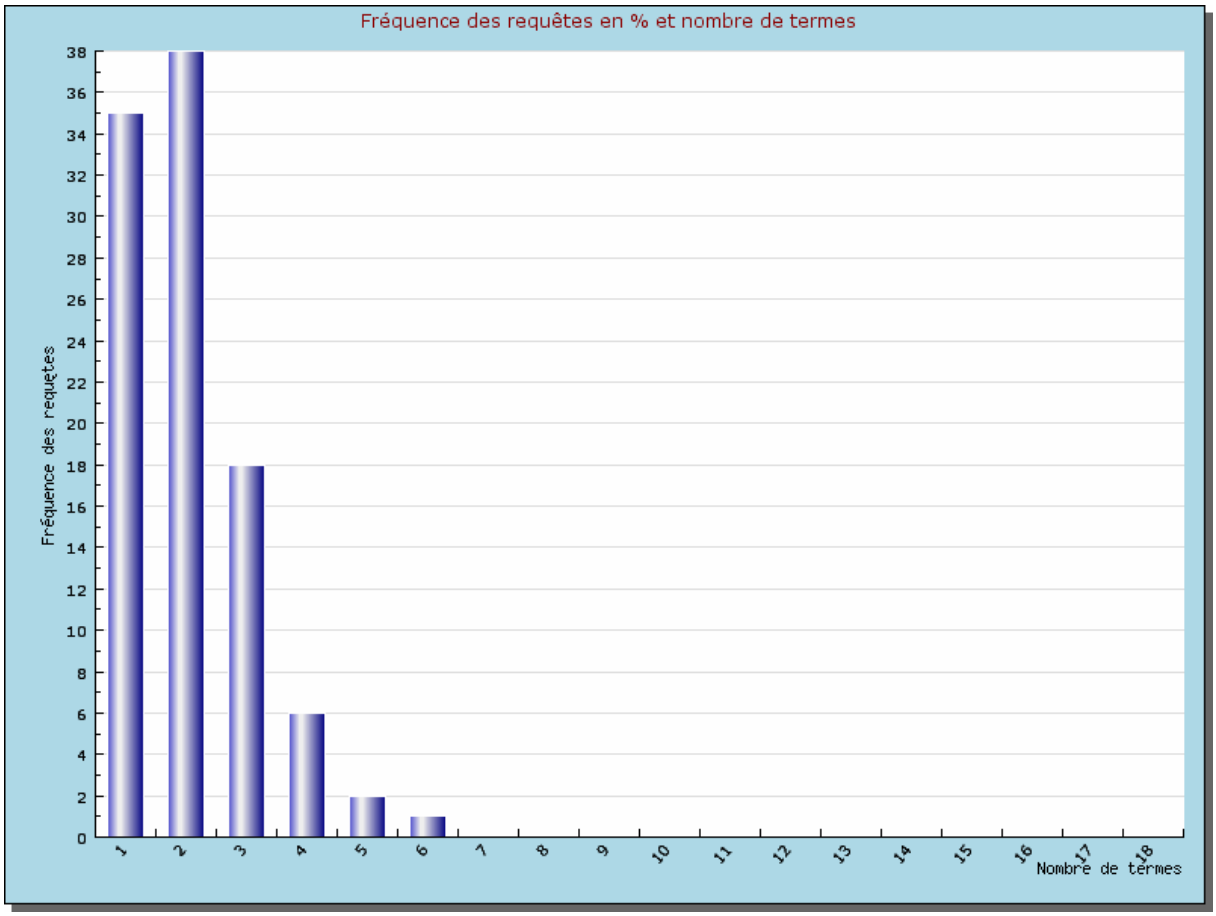

Figure 3. : Fréquence des requêtes en fonction du nombre de termes (Ensemble des requêtes)

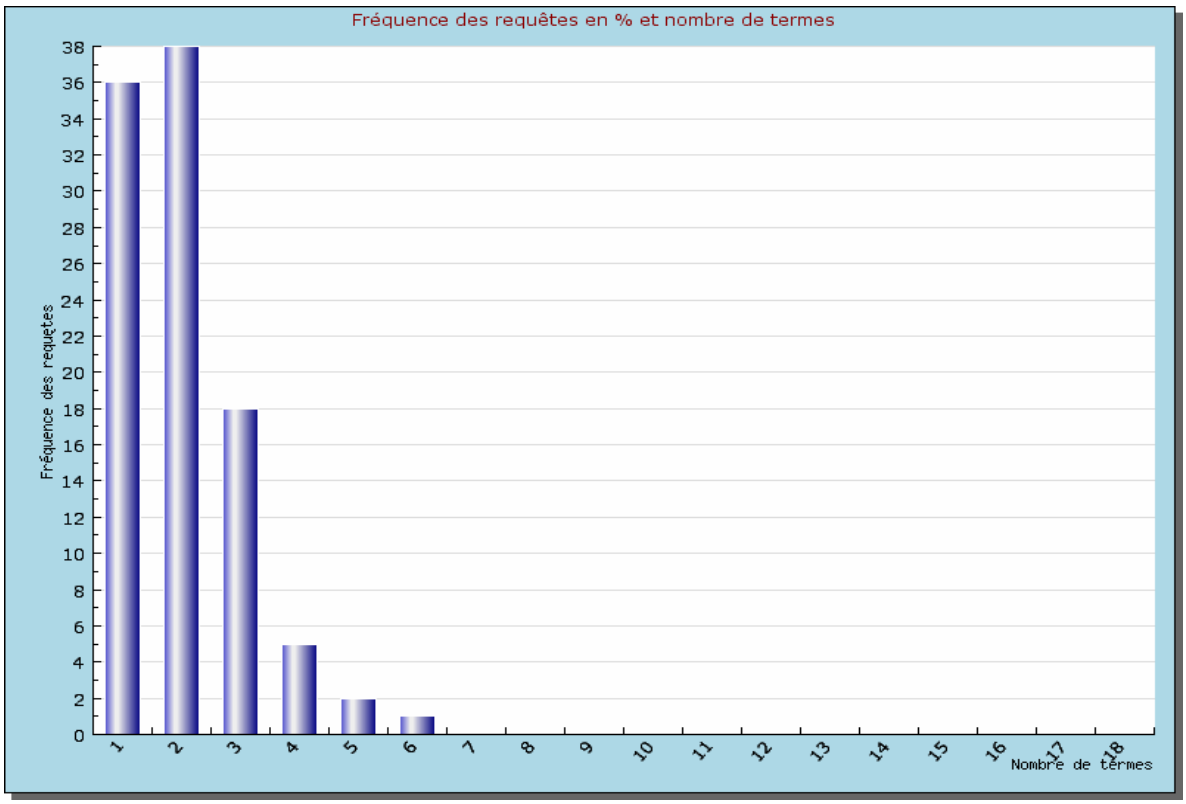

Figure 4. : Fréquence des requêtes en fonction du nombre de termes (Requêtes identifiées)

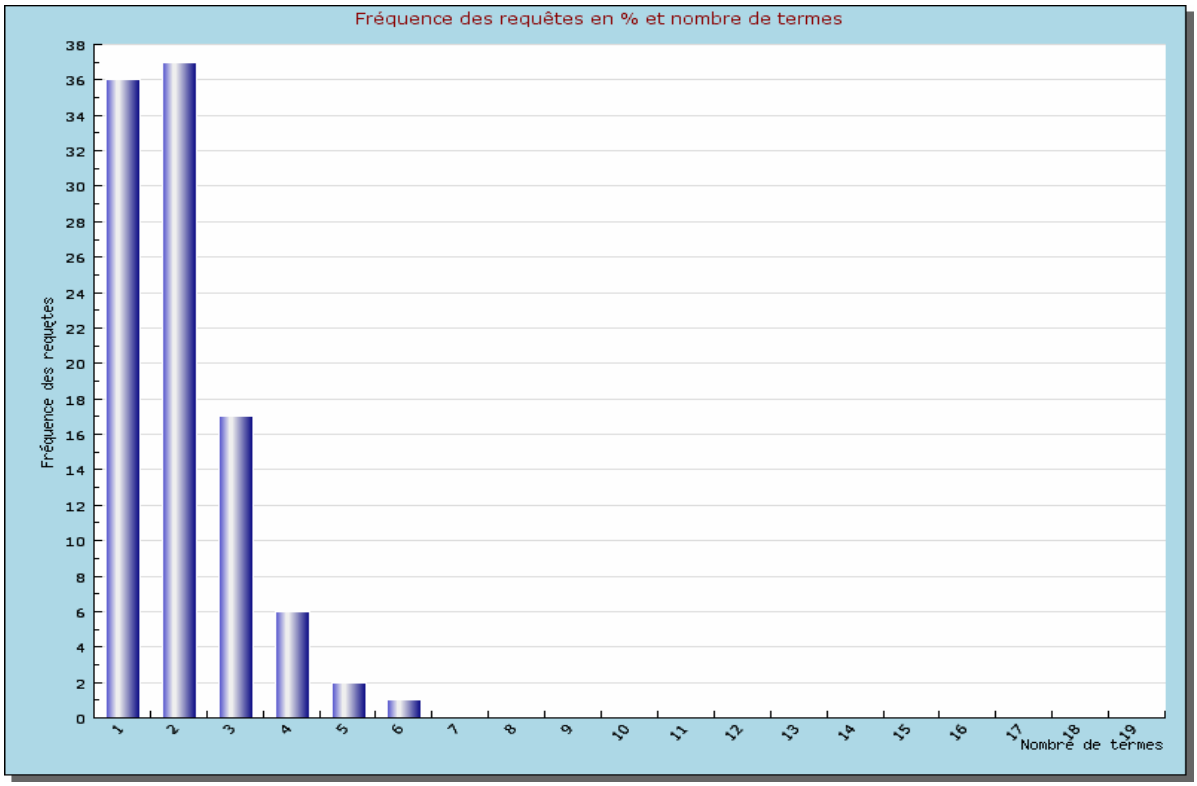

Figure 5. : Fréquence des requêtes en fonction du nombre de termes (Requêtes jamais identifiées)

On aurait pu penser que des utilisateurs *avisés*, c'est-à-dire sensibilisés au vocabulaire d'un domaine particulier, en l'occurrence celui des télécommunications, généreraient des requêtes *complexes*, c'est-à-dire contenant suffisamment de termes pour refléter au mieux leur interrogation et restreindre le nombre de résultats proposés. Or, ils semblent se comporter sur un moteur de recherche spécialisé (et de plus spécialisé dans ce qui devrait être leur domaine de compétence) comme les utilisateurs *grand public* se comportent sur les moteurs généralistes d'Internet. En effet une étude déjà ancienne (Spink, Bateman, et. Jansen, 1999) étudie le comportement de 18 000 internautes sur le moteur de recherche généraliste Excite. L'analyse des 51 000 requêtes montre que les requêtes à un terme représentent 31% de l'ensemble des requêtes, de même que les requêtes à deux termes. Les requêtes à trois termes représentent 18% des requêtes, et les requêtes à quatre termes ou plus sont rapidement décroissantes. On se rend donc compte que les utilisateurs de l'Aria n'ont pas un comportement très différent de celui décrit par Spink et al.

Utilisation des opérateurs booléens

Le moteur de recherche autorise l'utilisation des opérateurs booléens *AND* et *OR*, ainsi que l'opérateur de proximité *NEAR*. Deux types de guillemets sont autorisés. Une expression entrée entre doubles guillemets est recherchée telle qu'elle a été entrée, et une expression entrée entre simples guillemets est recherchée avec ses inflexions linguistiques.

Dans l'ensemble des requêtes

Sur les 46 210 requêtes différentes qui sont soumises au moteur, qu'elles aient suscité ou non un clic, 12 737 sont générées à l'aide d'opérateurs booléens, de proximité, et/ou de guillemets, soit environ 27%. Si l'on rentre dans le détail :

- 6 842 (14% des requêtes) sont générées au moins à l'aide d'un *AND*, *ET*, +
- 6 985 (15% des requêtes) sont générées au moins à l'aide de guillemets, doubles ou simples
- 1 882 (4% des requêtes) sont générées au moins à l'aide d'un *OR, OU*
- 42 sont générées à l'aide d'un *NOT*

- 2 utilisent au moins l'opérateur de proximité *NEAR*.

Dans les requêtes identifiées

Sur les 16 840 requêtes différentes qui sont soumises au moteur de recherche et qui ont suscité au moins un clic, 3 215 sont générées à l'aide d'opérateurs booléens, et/ou de guillemets, soit environ 19%. Si l'on entre dans le détail :

- 2 817 (17% des requêtes identifiées) sont générées au moins à l'aide d'un *AND*, *ET*, +
- 635 (3% des requêtes identifiées) sont générées au moins à l'aide de guillemets, doubles ou simples
- 797 (4% des requêtes identifiées) sont générées au moins à l'aide d'un *OR, OU*
- 13 sont générées à l'aide d'un *NOT*

La part de requêtes *complexes*, c'est-à-dire celles générées à l'aide d'opérateurs booléens ou de proximité, à l'aide de guillemets est donc plus élevée dans l'ensemble des requêtes que dans les requêtes identifiées. Ainsi, on peut en déduire que des requêtes complexes ne génèrent pas plus de clics que des requêtes générées sans opérateurs, alors qu'on aurait pu supposer le contraire, c'est-à-dire que des requêtes construites à l'aide d'opérateurs permettraient aux utilisateurs d'atteindre des documents reflétant mieux leurs demandes et leurs besoins.

Conclusion

Traditionnellement, dans le domaine de la recherche d'information, on suppose qu'une stratégie efficace est de générer des requêtes formées d'un certain nombre de mots afin de discriminer le nombre de termes. Or, si l'on regarde la façon dont sont construites les requêtes *efficaces* (c'est-à-dire celles qui suscitent au moins un clic), elles ne sont pas, au regard du nombre de termes qui les composent, différentes des requêtes *inefficaces* (c'est-à-dire celles qui ne suscitent pas de clic).

De plus, il semble que générer des requêtes à l'aide d'opérateurs booléens n'accroisse pas le taux de clic. En effet, 27% de l'ensemble des requêtes sont générées à l'aide d'opérateurs booléens ou d'opérateurs de proximité, alors que c'est le cas pour seulement 19% des requêtes ayant suscité au moins un clic. Ainsi, générer une requête avec des opérateurs booléens n'accroît pas la probabilité de réussite de la requête.

4. Les enseignements et les changements induits par cette étude

L'analyse du comportement des utilisateurs de l'Arianet face au moteur de recherche *Verity Search'97* va nous amener à proposer des outils et des produits d'information destinés à les aider à trouver l'information dont ils ont besoin, et qu'ils n'arrivent pas à trouver sur le serveur par leurs propres moyens.

Le nombre de visiteurs

Dans une première analyse, le nombre moyen d'utilisateurs différents identifiés sur le serveur via le moteur de recherche, c'est-à-dire accédant à des documents en effectuant une requête, est de 538. Or, en moyenne, 2 000 visiteurs différents s'identifient chaque mois sur le site. Ainsi chaque mois, un tiers seulement des visiteurs vont utiliser avec succès le moteur de recherche pour accéder à des documents. On peut ainsi en déduire qu'ils valorisent au moins tout autant la navigation par source, ou bien la catégorisation thématique proposée sur la page d'accueil du site.

Le nombre de termes constituant les requêtes

Les utilisateurs de l'Arianet construisent leurs requêtes en moyenne à l'aide de 2,13 termes. Le mode est constitué de deux termes, c'est-à-dire que ce sont les requêtes à deux termes qui sont, en valeurs absolues, les plus fréquentes (très proches des requêtes à un terme). Par ailleurs, les requêtes à un ou deux termes constituent presque les trois quarts des requêtes.

Comme ces données se vérifient quel que soit le type des requêtes étudiées (ensemble des requêtes, requêtes identifiées, ou requêtes non identifiées), on peut étonnamment en déduire que dans les données étudiées, le nombre de termes constituant les requêtes n'influence pas le *taux de clic*. Malgré ce que l'on pourrait supposer, des requêtes constituées d'un nombre relativement élevé de termes ne sont pas plus efficaces que des requêtes à un ou deux termes.

Perspectives

Afin d'accroître le système d'information, c'est-à-dire dans notre cas d'améliorer la qualité des réponses du moteur de recherche, les utilisateurs ont besoin d'autres outils et d'autres types de représentation pour accéder à l'information, tels que :

- les outils de catégorisation ;
- les outils d'extraction ;
- les newsletters et les mécanismes d'alertes pour des objets identifiés.