Le chercheur peut-il se fier aux volumétries indiquées par les moteurs de recherche commerciaux ?

Résumé - Les moteurs de recherche commerciaux tels que Google, Bing ou Yahoo! ont séduit les chercheurs de diverses disciplines pour l'utilisation des résultats de recherche ou du nombre de résultats estimé. La webométrie est la discipline dédiée à l'étude de ces pratiques. De nombreuses études montrent les limites des moteurs en la matière. Ces études deviennent cependant rapidement dépassées du fait des modifications régulièrement apportées aux techniques d'indexation, à la syntaxe d'interrogation ou aux interfaces de programmation (API). Cette recherche s'attachera à confirmer les limites mises en évidence pour les moteurs de recherche Bing et Google, et identifiera de nouvelles limitations liées à l'utilisation des opérateurs booléens.

Mots-clefs: moteur de recherche, bing, google, webométrie, api.
Axe thématique: « Pratiques numériques », « Les modalités de l'usage des TIC : ergonomie, acceptabilité, usages ».

Contexte

La consultation de la messagerie et la recherche d'information restent les deux premiers cas d'utilisation d'Internet (Berret, 2008). Trémenbert (2010) les estime respectivement à 80% et 92% (tout en notant de fortes disparités d'usage). Les chercheurs font également un large usage des moteurs de recherche publics, que ce soit pour leurs recherches d'information, ou pour exploiter les listes de résultats et les informations associées.

L'étude quantitative des phénomènes relatifs au Web fait l'objet d'une discipline particulière: la webométrie (Thelwall et al., 2005). Thelwall, Vaughan et Björneborn (2005) notent notamment un usage répandu des décomptes de pages de résultats. Les principaux moteurs de recherche publics (Google, Yahoo!, Bing) donnent en effet accès à des index de très grande taille. Sous une apparence séduisante d'exhaustivité, les auteurs insistent cependant sur la couverture partielle du Web par ces outils ainsi que sur les biais potentiels introduits par les disparités nationales ou les secrets commerciaux. Une étude des résultats menée par Véronis (2006) montre par ailleurs l'apparition d'associations préférentielles entre sites commerciaux et moteurs de recherche, parfois explicables par l'existence de partenariats connus du public.

Notre recherche concerne les implications pour le chercheur de l'exploitation des volumétries (« *hit count* ») des résultats de recherche renvoyées par les moteurs de recherche. Nous identifierons en particulier un ensemble de points susceptibles d'affecter les conclusions de recherches recourant à ce type de données.

Revue de la littérature

Les cas d'utilisation de la volumétrie des listes de résultats sont diversifiés. Ils concernent notamment les domaines du traitement de la langue (Cimiano et al., 2003; Grefenstette, 1999; Turney, 2001), l'analyse de sentiments dans les textes (Turney, 2002), l'analyse des performances des entreprises (Romero-Frias, 2009), la diffusion des publications académiques auprès des entreprises (Thelwall, 2004), ou encore l'évaluation automatique d'articles ou de revues scientifiques (Thelwall, 2004; Chena et al., 2006; Moussaa et Touzania, 2009).

Les moteurs de recherche interdisent souvent l'extraction de données par l'exécution de requêtes automatiques (McCown et Nelson, 2007a, 2007b). Les moteurs de recherche de Google et de Microsoft disposent d'ailleurs de mécanismes pour détecter et bloquer l'exécution de requêtes automatiques (Thelwall et Sud, 2012). En contrepartie, les moteurs de recherche les plus importants, tels Google, Microsoft ou Yahoo!, proposent des API (*Application Programming Interface*), en complément de la WUI (*Web User Interface*), permettant à un logiciel externe d'automatiquement interroger le moteur et d'en exploiter les réponses. Des moteurs de recherche plus modestes proposent des services similaires. C'est le cas pour Amazon A9, Entireweb Search, Naver, Yandex ou Gigablast (Thelwall et Sud, 2012).

Les API des trois principaux moteurs de recherche commerciaux (et, plus largement, les moteurs eux-mêmes) ont fait l'objet de nombreuses évolutions depuis leur création. Ces évolutions concernent les technologies mises en œuvre pour l'indexation des textes, la syntaxe d'interrogation proposée, les limitations contractuelles à l'utilisation des API, etc.

L'étude du fonctionnement des API permet par ailleurs de constater des comportements nuisant à la fiabilité des résultats obtenus. Mayr et Tosques (2005) ont ainsi comparé les résultats provenant de l'API Google (première version) et de la WUI. Ils ont constaté de fortes différences dans le classement des pages Web mais aussi dans le nombre de résultats renvoyés par requête. McCown et Nelson (2007a, 2007b) ont comparé le fonctionnement des API de Google, de Yahoo et de MSN (premières versions). Le nombre de résultats issus de MSN est apparu sensiblement le même pour l'API et la WUI, au contraire de Yahoo! et de Google. Par contre, le nombre de pages indexées par site ou de *backlinks* différaient sensiblement, pour tous les moteurs, entre l'API et la WUI. Ils suggèrent que Google et Yahoo! s'appuient sur un index plus petit pour son API (mais pas plus ancien). Funahashi et Yamana (2010) ont pour leur part confirmé l'instabilité des résultats dans le temps.

Compte tenu de l'évolution rapide des moteurs et de leurs API, la plupart des études techniques sont aujourd'hui dépassées. Ainsi, McCown et Nelson (2007a, 2007b) s'appuient sur la première version de l'API Google, sur l'API MSN et l'API REST de Yahoo!. Toutes ont été arrêtées depuis, et remplacées par une nouvelle version. Leur fonctionnement a dès lors pu être amélioré ou dégradé. Par ailleurs, certains facteurs n'ont pas été pris en compte, tel que le comportement en fonction du type de requête -requête simple, composée ou complexe (Shafi et Rather, 2005)- ou la qualité des ciblages géographiques.

Hypothèses

Nous proposons de tester les trois hypothèses suivantes:

1. Les requêtes complexes, ou booléennes, donnent des résultats conformes à la théorie des ensembles.
2. La volumétrie donnée par l'API est différente de la volumétrie donnée par la WUI.
3. La volumétrie donnée par l'API n'est pas proportionnelle à la

volumétrie donnée par la WUI.

Méthodologie

Cette recherche a été initiée dans la cadre d'une étude sur la popularité des marques automobiles sur Internet. Dans ce cadre, des incohérences entre résultats de recherche ont été découvertes. Leur existence a été confirmée par un état de l'art sur le sujet. Nous réaliserons donc notre expérimentations en utilisant comme mots-clefs des noms de marques ou de modèles.

En pratique, nous avons créé une liste de 20 couples marque-modèle. Nous les avons ensuite traité pour donner les requêtes suivantes: 'marque', 'modèle', 'marque modèle', 'marque -modèle', 'marque AND modèle' et 'marque OR modèle', soient 120 requêtes. Ces requêtes ont été exécutées sur Bing (WUI et API) ainsi que sur Google (WUI et API) avec les paramètres par défaut. Elles s'appliquent à la totalité de l'index (Web mondial). L'ancienne API de Google a été utilisée du fait des restrictions d'utilisation associées à la nouvelle version (maximum 100 requêtes gratuites par jour).

Les données collectées sont automatiquement mises en forme dans un tableau, exploitable dans un tableur. Elles sont associées à la date d'exécution et aux URLs générées pour interroger les moteurs. Des vérifications peuvent ainsi être faites quant au bon déroulement de la collecte de données.

La première hypothèse sera testée en calculant les rapports suivants:

$$q1 = \frac{card('marque\,modele') + card('marque\ -modele)}{card('marque')}$$

$$q2 = \frac{card('marque\ \mathrm{AND}\ modèle') + card(marque\ \mathrm{OR}\ modèle)}{card('marque') + card('modele')}$$

La seconde hypothèse sera testée en considérant le rapport de volumétrie entre chaque moteur et son API.

La troisième hypothèse sera testée en considérant la corrélation entre la volumétrie renvoyée par l'interface utilisateur de chaque moteur et son API.

Résultats

Hypothèse 1:

Si les moteurs de recherche respectent la logique booléenne, les rapports `q1` et `q2` doivent être proches de 1.

	q1 (moyenne)	**q2 (moyenne)**
Bing	0.98	1.00
Google	3.43	1.95

Table 1 - Rapports entre nombres de résultats obtenus et de résultats attendus.

En pratique, si le fonctionnement de Bing paraît conforme au comportement attendu, il n'en est pas de même pour Google. Pour ce dernier, la valeur du rapport `q1` « s'explique » par le résultat renvoyé par la requête de type 'marque -modèle', indiquant systématiquement plus de résultats que la requête 'marque'.

Nous constatons par ailleurs que les deux moteurs sont sensibles à la casse pour l'opérateur

« OR ». La même requête avec « or » ou « OR » sera donc interprétée différemment. De même, une requête à deux termes avec ou sans « AND » donnera des résultats différents chez Google.

Hypothèse 2:

Dans le cas de Google, le nombre de résultats estimés d'une requête sur l'interface utilisateur du moteur de recherche diffère de ce même nombre de résultats estimé par l'interface de programmation. Le nombre de résultats renvoyé par Google est également plus élevé que celui renvoyé par Bing. L'API et la WUI de Bing renvoient par contre des résultats fort proches.

Geng/Gapi	17.38
Beng/Bapi	1.04
Geng/Beng	4.18

Table 2 – Rapports entre nombres de résultats estimés. Légende : B = Bing, G = Google, eng = WUI, api = API.

La différence entre WUI et API pour Google pourrait s'expliquer par la taille de l'index sous-jacent (McCown et Nelson (2007a, 2007b). La différence entre Bing et Google pourrait s'expliquer par la même raison, ou par la manière de comptabiliser ou non les pages similaires ou dupliquées. Google gère au moins 1 billion de liens mais son index ne fait « que » quelques milliards de pages (Google, 2008; Boughanem et al., 2006).

Les estimations données par l'API et la WUI de Bing ont dans un premier temps différé, avant de devenir fort proches. Les premières expérimentations ont peut-être été réalisées lors d'une mise à jour de l'index utilisé par l'API, comparable aux « danses » connues chez Google voici quelques années.

Hypothèse 3:

La corrélation entre la WUI de Google et son API est élevée, de même que la corrélation entre les WUI de Bing et de Google. Cependant, elles se dégradent lorsque l'on prend en compte uniquement les requêtes simples. Ces valeurs fluctuent par ailleurs avec le temps dans le cas des requêtes simples. Les volumétries estimées par la WUI et l'API de Bing sont par contre équivalentes dans tous les cas testés (après avoir, dans un premier temps, donné des résultats différents).

	Toutes les requêtes	**Requêtes simples uniquement**
Beng, Bapi	1.00	1.00
Geng, Gapi	0.93	0.71
Beng, Geng	0.81	0.66

Table 3 - Corrélations. Légende : B = Bing, G = Google., eng = WUI, api = API.

Conclusion

Bien que la volumétrie des résultats des moteurs de recherche commerciaux (ou « *hit counts* ») soit discutée depuis plusieurs années, les études pratiques ne sont pas fréquentes. De plus, leurs résultats sont rapidement dépassés. Les moteurs de recherche commerciaux évoluent en effet en permanence, que ce soit au niveau de leur algorithme de recherche, de leur syntaxe ou des interfaces de programmation (API) permettant automatiquement d'accéder aux jeux de résultats.

Les expérimentations ont permis de confirmer que la volumétrie indiquée par les interfaces de programmation d'applications (API) peut différer sensiblement de celle indiquée par les interfaces Web pour utilisateur (WUI). Les expérimentations ont par ailleurs permis d'identifier de nouveaux

problèmes dans les volumétries renvoyées par les requêtes utilisant implicitement (requête composée de deux termes) ou explicitement (utilisation des opérateurs OR et AND) des opérateurs booléens. Enfin, des variations de comportement ont été observées pour Bing durant la période de test. Elles pourraient provenir d'une mise à jour de l'index de l'API comparable aux « danses » connues avec le moteur Google voici quelques années.

Le comportement de Bing apparaît plus stable et prévisible que celui de Google. Un usage prudent des valeurs renvoyées par Bing paraît donc envisageable, en particulier dès lors que les proportions intéressent plus le chercheur que la précision des valeurs en elles-mêmes. Nous recommandons cependant une observation attentive des valeurs renvoyées avant de les utiliser pour une application pratique.

Des tests supplémentaires sont en cours de réalisation. Ils concernent notamment le comportement de la dernière version de l'API de Google ainsi que l'influence du ciblage géographique sur les volumétries estimées par les moteurs.

Bibliographie

Berret P. (2008). Diffusion et utilisation des TIC en France et en Europe. Culture chiffres, 2008/2 n°2, pp. 1-15.

Boughanem M., Tamine-Lechani L., Martinez J., Calabretto S., Chevallet J.-P. (2006). Un nouveau passage à l'échelle en recherche d'information. Ingénierie des Systèmes d'Information (ISI) 11, 4, pp. 9-35.

Chena P., Xieb H., Maslovc S., Rednera S. (2007). Finding scientific gems with Google's PageRank algorithm. Journal of Informetrics 1, pp. 8–15.

Cimiano P., Pivk A., Schmidt-Thieme L., Staab S. (2003). Learning Taxonomic Relations from Heterogeneous Sources of Evidence. IOS Press.

Funahashi T., Yamana H. (2010). Reliability Verification of Search Engines' Hit Counts: How to Select a Reliable Hit Count for a Query. Computer Science, 2010, Volume 6385/2010, pp. 114-125.

Google (2008). We knew the web was big.... 25 juillet 2008. Site: googleblog.blogspot.com (consulté le 25/01/2012).

Grefenstette G. (1999). The World Wide Web as a resource for example-based machine translation tasks. Translating and the Computer (Proceedings), November 10-11, 1999.

Kilgarriff A. (2007). Googleology is Bad Science. Computational Linguistics 33 (1), pp. 147-151.

Mayr P., Tosques F. (2005). Google Web APIs - an instrument for Webometric analyses?. Proceedings of the ISSI 2005 conference.

McCown F., Nelson M.L. (2007a). Search engines and their public interfaces: which apis are the most synchronized?. Proceedings of the 16th international conference on World Wide Web.

McCown F., Nelson M.L. (2007b). Agreeing to disagree: search engine and their public interface. JCDL'07, June 18-23, 2007.

Moussa S., Touzani M. (2010). Ranking marketing journals using the Google Scholar-based hg-index. Journal of Informetrics 4, pp. 107–117.

Romero-Frias E. (2009). Googling Companies - A Webometric Approach to Business Studies. Electronic Journal of Business Research Methods, 7(1), pp. 93-106.

Shafi S.M., Rather R.A. (2005). Precision and recall of five search engines for retrieval of scholarly information in the field of biotechnology. Webology, Volume 2, Number 2, August 2005.

Taboada M., Anthony C., Voll K. (2006). Methods for creating semantic orientation dictionaries. Proceedings of Fifth International Conference on Language Resources and Evaluation (LREC 2006). Genoa, Italy. pp. 427-432.

Thelwall M. (2004). Can the web give useful information about commercial uses of scientific research? Online Information Review, 28(2), pp. 120-130.

Thelwall M., Vaughan L., Björneborn L. (2005). Webometrics. In: Annual Review of Information Science and Technology 39, pp. 81-135.

Thelwall M., Sud, P. (2012). Webometric research with the Bing Search API 2.0. Journal of Informetrics, 6(1), pp44-52.

Turney P.D. (2001). Mining the Web for synonyms: PMI-IR versus LSA on TOEFL, Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001), Freiburg, Germany, pp. 491-502.

Turney P.D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02), Philadelphia, Pennsylvania, pp. 417-424

Trémenbert J. (2010). Point sur les usages d'Internet: usage des réseaux sociaux et e-participation. 12 octobre 2010. Site: marsouin.org (consulté le 16/01/2012).

Ulyar A. (2009). Investigation of the accuracy of search engine hit counts. Journal of Information Science, August 2009, vol. 35, n°4, pp. 469-480.

Véronis J. (2006). Étude comparative de six moteurs de recherche. Université de Provence, 23 février 2006.

***