Résumé Article
« Can Social Bookmarking Improve Web Search? »
Source : http://ilpubs.stanford.edu:8090/817/1/2007-33.pdf

L'article s'intitule « Can Social Bookmarking Improve Web Search? », il a été publié en Novembre 2007 dans « Infolab Technical Report » et il a été écrit par Paul Heymann. Cet article présente une étude expérimentale qui a été réalisée sur un des sites majeurs de Social Bookmarking : « del.icio.us ».

Le Social Bookmarking est un phénomène permettant de fournir un certain nombre de données sur des pages web, il s'agit d'une forme différente de site social impliquant des signets partagés tel que les réseaux sociaux liés à l'actualité. Dans le Social Bookmarking, il existe 3 types de données : la structure de lien, le contenu de la page et le contenu décrivant la page (tag // signet) généré par l'utilisateur.

Le but de l'article est de quantifier la taille de ce dernier type de donnée, de caractériser les informations qu'il contient et de déterminer l'impact qu'il peut avoir sur l'amélioration de la recherche sur le web.

Le Social Bookmarking inclus 3 unités de données : triple, post et label. Le Triple de la forme <utilisateur i, tag j, url k> est un tuple signifiant que l'utilisateur i a marqué l'url k avec la balise j. Le Post est un url marquée par un utilisateur et toutes les métadonnées associées. Un post est composé de plusieurs triples, mais il peut également contenir des informations comme un commentaire de l'utilisateur. Le Label de la forme <tag i, url k> est une paire qui signifie qu'au moins un triple contenant une balise i et un url k existe dans le système.

L'article se consacre au site del.icio.us car il s'agit d'un des sites principaux du Social Bookmarking. En général, les entreprises qui contrôlent les sites sociaux effectuent un certains nombre d'analyses internes, mais elles ne publient pas de résultats spécifiques, afin d'assurer la confidentialité des utilisateurs.

Del.icio.us offre une variété d'interfaces selon différentes parties. On retrouve un flux récent qui fournit les signets les plus récents publiés en temps réel mais seulement certains posts sont publiés en raison du filtrage. Il existe également des interfaces qui montrent tous les messages concernant une url donnée, tous les messages d'un utilisateur donné et tous les messages les plus récents pour une balise donnée. Cependant, l'interface «messages pour un utilisateur donné» n'est pas filtrée, car les utilisateurs partagent souvent cette interface avec d'autres utilisateurs pour leur donner une idée de leurs signets actuels.

Ces différentes interfaces permettent deux stratégies différentes dans la collecte des données : la surveillance de l'actualité et le traitement du site comme étant un graphe triparti. Ces deux stratégies se complètent. La surveillance est biaisée par rapport aux pages populaires, tandis que l'exploration tend à être biaisée vers ces pages.

L'étude s'est portée sur 3 jeux de données différents : à grande échelle (C(rawl)), selon les derniers posts (R(ecent)), et au courant d'un mois (M(onth)). Les données du site sont importantes et croissent rapidement. Certains messages, utilisateurs ou tags peuvent être manquants en raison du filtrage ou du processus d'analyse. Enfin, les données peuvent être biaisées peuvent être sur-représentées.

Pour plus de 2 000 utilisateurs échantillonnés au hasard, il y a deux conclusions.

Tout d'abord, en moyenne, environ 20% des messages publics n'apparaissent pas dans le flux récent. Deuxièmement, les url populaires, les url de domaines populaires, les messages utilisant des méthodes automatisées et le spam ne figurent souvent pas dans le flux récent. Les messages manquants dans l'ensemble se réfèrent à des URL visiblement plus populaires, mais l'effet de leur absence semble minime.

Depuis le début du web, les gens ont utilisé le contenu de la page pour faciliter la navigation et la recherche. En 1994 les utilisateurs suggéraient l'utilisation d'un texte d'ancrage et d'une structure de liens pour améliorer la recherche sur le Web. Pendant ce temps, il y a eu aussi un courant d'utilisateurs qui tentait d'annoter leurs propres pages avec des métadonnées. Cela a commencé avec la balise <meta> qui permettait aux mots-clés sur une page Web d'aider les moteurs de

recherche. Toutefois, en raison du spam des moteurs de recherche, cette pratique a disparu.. Indépendamment de la recherche sur le Web, le marquage a suscité un intérêt croissant. Ceci est principalement dû à son utilité comme outil d'organisation léger et comme un moyen d'augmenter le texte pour la recherche de vidéo et d'image.
On constate certains facteurs positifs dont les signets qui permettent à une personne de se souvenir des URL visitées, mais également les étiquettes qui peuvent être faites par la communauté pour guider les utilisateurs vers un contenu recherché.
Concernant les url, l'étude a montré que les pages publiées sur le site sont souvent modifiées et les utilisateurs consultent surtout ces pages car elles sont activement mises à jour ou qu'elles viennent d'être crées. Les résultats ont également montré qu'environ 12,5% des url présentes sont de nouvelles pages non indexées. Les 4 majeures causes sont : les pages sont indexées sous une autre url, elles sont de type image par exemple, elles n'existent pas, ou il s'agit d'une page de spam. Le Social Bookmarking semble être une bonne source de nouvelles pages actives, il aide les moteurs de recherches à découvrir des pages. On trouve également que 9% des résultats pour les requêtes de recherche sont des url présentes dans del.icio.us, ce qui montre que les url ne sont pas proportionnellement courante dans les résultats de recherche par rapport à leur couverture.
Concernant les tags, les résultats de la recherche a montré que les termes de requête et les balises populaires se chevauchent de manière significative (bien que les balises et les termes de requête ne soient pas corrélés). Mais également que la plupart des étiquettes ont été jugées pertinentes et objectives par les utilisateurs. L'étude a pu conclure que les tages sont dans l'ensemble assez précis.