Quelle est l'influence de la taille d'un site dans une stratégie SEO ?

Retour au sommaire de la lettre

Domaine :	Recherche	**Référencement**
Niveau :	Pour tous	**Avancé**

Une récente vidéo de Matt Cutts évoquait la question de la taille d'un site web (le nombre de ses pages indexées par le moteur) dans le cadre du SEO. Sa réponse, un peu confuse, n'allait, comme d'habitude, pas assez loin dans cette analyse. Car, en effet, la taille d'un site a clairement un poids dans le trafic SEO généré. Même si, depuis la mise à jour de l'algorithme Mayday et le filtre Panda, cet avantage a diminué au fil des ans, il reste néanmoins non négligeable. Exploration...

Nota : Les URL correspondant aux sources citées dans cet article se trouvent à la fin.

"Plus un site est gros, mieux ses pages seront référencées...". Cette affirmation revient régulièrement dans la bouche de certains référenceurs. Certains affirment même parfois que c'est un critère dans l'algorithme de Google. Barry Schwartz (l'éditeur de Seroundtable.com) a eu l'occasion de faire un sondage parmi ses lecteurs en juillet 2013, et 70% de ceux qui ont répondu pensent que la taille d'un site influe sur son référencement.

Récemment, ce sujet a été débattu sur le forum américain WebmasterWorld, sur le site Moz.com, et il vient de faire l'objet d'une vidéo de Matt Cutts en octobre dernier.
Mais qu'en est-il vraiment ? Est-ce que la taille d'un site a une réelle influence sur le référencement ? Nous allons voir que la réponse est oui, en théorie, mais qu'en pratique, ce n'est pas aussi simple, et que l'influence de la taille du site a beaucoup diminué en quinze ans ...

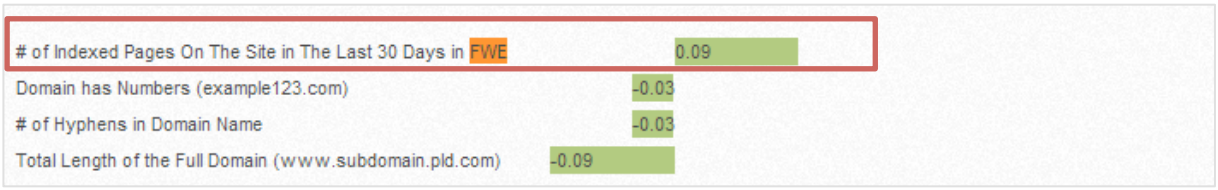
Pas de corrélation claire entre le positionnement d'une page et la taille du site qui la contient ?

Depuis quinze ans que des observateurs cherchent à faire de l'ingéniérie inversée sur l'algorithme de Google, un faible nombre d'études ont été réalisées pour démontrer l'existence d'une corrélation entre taille du site et positions.

On trouve néanmoins quelques tentatives, et le moins que l'on puisse dire, c'est que les résultats ne sont en général pas très concluants.

L'étude de SEOMoz

L'équipe de Rand Fishkin (Moz.com) fait tous les ans une étude (http://moz.com/search-ranking-factors) pour savoir quels sont les facteurs de classement que les référenceurs pensent être à l'oeuvre dans l'algorithme de Google. Le nombre de pages indexées fait évidemment partie des critères évoqués par les SEOs interrogés. Mais si l'on regarde les résultats de l'étude statistique, on ne peut pas dire que les résultats soient très probants :



Un coefficient de Pearson de 0,09 indique une très faible corrélation. *A contrario*, on peut même dire que très probablement la position d'une page sur une requête donnée est très peu influencée par la taille du site qui contient cette page.

L'étude de Website Magazine

Une expérience similaire (mais un peu moins "scientifique") a été tentée en 2010 par le site Website Magazine. On retrouve bien les mêmes résultats que ceux de SEO Moz : si on met en évidence la taille des sites en regard des positions sur une page de résultats, on en tire assez vite la conclusion qu'un "petit" site a autant de chances qu'un "gros" de sortir sur les requêtes testées.

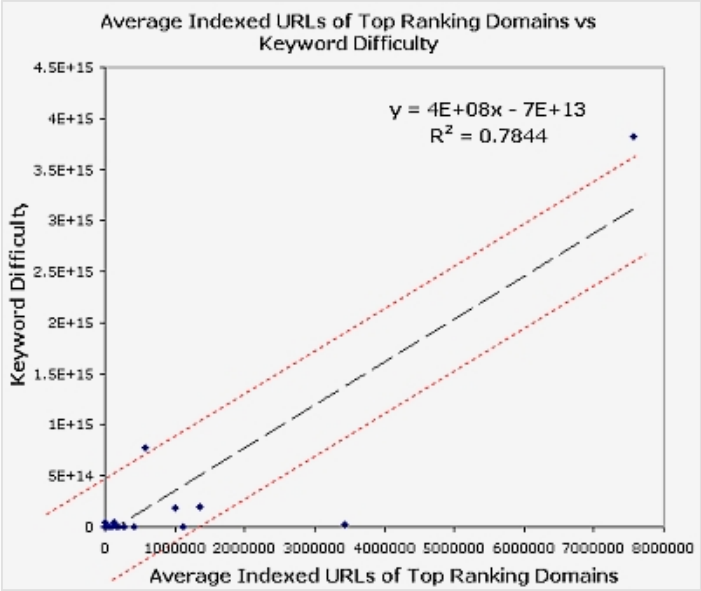
	KEYWORD: FREE RINGTONES @ GOOGLE	Google	Google
	URL	with WWW	w/o WWW
1	www.myxer	661,000	978,000
2	www.phonezoo.com	115,000	115,000
3	www.mytinyphone.com	71,000	71,000
4	www.eztracks.com		
5	www.pisamba.com	65,000	65,000
6	www.brinked.com	53,000	80,000
7	www.tone9.com		
8	www.funformobile.com	2,400	154,000
9	www.ventones.com	6,600	6,600
10	www.agorics.com	37,000	37,000
11	www.mobiletrend.net	452	452
12	www.coolfreeringtones.com	2,400	2,400
13	www.zedge.net	4,200,000	5,400,000
14	www.mobiles24.com	257,000	276,000
15	www.mobile17.com		
16	www.ringtones4all.com	1,400	1,500
17	www.mymojo.com	445	563
18	www.cellfish.com	2	41,000
19	www.ringophone.com	15,000	31,000
20	www.free-ringtones-download.ws		

Un exemple de requête étudiée dans l'étude de website magazine : sur la requête « free ringtones », la taille du site ne semble pas influencer les classements de manière convaincante.

Mais est-ce vrai sur toutes les requêtes ?

L'étude de SEO Chat de 2010

Jim Boykin a tenté, dans une étude publiée sur SEO Chat en 2010, de regarder s'il y'avait une corrélation entre les chances de se positionner correctement sur des requêtes concurrentielles, et la taille du site.



Exemple de graphe tiré de l'étude pour analyser la corrélation entre les deux variables.

Ses conclusions sont claires : il existe une corrélation. Mais il ajoute également - et prudemment - qu'il y a probablement des corrélations croisées entre la taille du site et d'autres facteurs. C'est à dire que les vraies "causes" proviennent peut-être d'autres signaux qui sont plus "forts" sur de gros sites que sur des petits.

Une corrélation logique : taille du site et trafic généré

Une étude réalisé par des spécialistes de la bibliométrie, et publiée dans la revue "Hi Tech Library" a cherché à établir les facteurs déterminant le trafic d'un site de bibliothèque en ligne. La corrélation la plus forte a été trouvée avec... le nombre de pages indexées dans les moteurs de recherche.

Mais cette corrélation est logique : la "visibilité" d'un site est forcément en relation avec le nombre d'expressions sur lesquelles ce site se positionne. Et ce nombre d'expressions est en relation étroite avec... le nombre de pages indexées par les moteurs pour ce site.

Table VI.
Spearman rho correlation test for daily page views and number of Google indexed pages

	Spearman rho correlation Website total daily page views	Website total Google indexed pages
Website total daily page views		
Pearson correlation	1	0.639*
Sig. (two-tailed)		0.000
Website total Google indexed pages		
n	52	52
Pearson correlation	0.639**	1
Sig. (two-tailed)	0.000	
n	52	52

Notes: *Correlation is significant at the 0.01 level (two-tailed)

Un extrait de l'étude de la revue « Hi Tech Library »

Plus un site est "gros", plus il capte de trafic en provenance des moteurs de recherche

C'est ce qu'explique clairement Matt Cutts dans sa vidéo publiée le 28 octobre 2013 : plus un site comporte de pages, plus il a d'opportunités de se positionner sur un nombre plus étendu de requêtes. C'est aussi simple que cela !

Une capture d'écran de la dernière vidéo de Matt Cutts sur le sujet, publiée le 28 octobre 2013

On tient là la principale raison de cette influence de la taille du site sur le référencement... Mais Matt Cutts insiste ensuite pour expliquer qu'il existe un deuxième élément favorisant les gros sites : avoir beaucoup de pages maximise vos chances de recevoir des backlinks... Bref un gros site récupère plus facilement du pagerank qu'un petit site. Matt Cutts donne donc les deux raisons qui favorisent les gros sites : ils ont une plus grande "surface sémantique", et de meilleures chances d'avoir un pagerank élevé.

Remarquons que la position de Matt Cutts sur ce sujet est devenue plus claire et moins dans le déni que dans sa précédente vidéo sur le sujet (datée d'octobre 2009), dans laquelle il répondait que la taille du site avait zéro influence sur les classements. Mais la question était différente : "*Does the size of a website affect its authority in Google?*".

Mais Matt Cutts ne dit pas tout dans sa réponse... Il oublie de préciser le rôle du linking interne dans l'algorithme de Google notamment dans sa diffusion du Pagerank. Or le maillage interne favorise théoriquement de manière spectaculaire les gros sites.

Les gros sites partent avec un bonus de PageRank

L'analyse mathématique de l'algorithme du pagerank réserve quelques surprises. On découvre notamment qu'un site qui ne reçoit aucun lien entrant (aucun "backlink") peut néanmoins avoir un pagerank !

Une page qui ne comporte pas de liens a quand même un pagerank...

$$PR(A) = \frac{1-d}{N} + d\left(\frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)} + \cdots\right)$$

La formule du Pagerank (telle que présentée dans le deuxième article de Page & Brin)

Tout d'abord, en regardant la formule, on voit au premier coup d'oeil qu'il existe deux termes dans la formule, un qui dépend des pagerank transmis par les backlinks (divisé par le nombre de liens sortant, et un facteur "(1-d)/N".

- N ici représente le nombre de pages total dans l'index.

- d représente le facteur d'atténuation (d = "damping factor"), un coefficient qui empêche 100% du pagerank d'être transmis aux pages suivantes par les liens. Dans la formule initiale du pagerank (celle mentionnée dans l'article séminal de Page et Brin), ce facteur avait une valeur de 0,85.

Conclusion : une page isolée a forcément un PR de (1-0,85)/1 = 0,15 , puisque que le deuxième terme à une valeur nulle, faute de backlinks !

Un site qui ne reçoit pas de backlinks peut concentrer un pagerank important

Trois chercheurs italiens de l'Université de Sienne en Italie ont publié en 2004 un article passionnant sur les propriétés de l'algorithme du Pagerank : "Inside Pagerank" (Monica Bianchini, Marco Gori et Franco Scarselli).

Ils ont en particulier cherché à comprendre la manière dont le pagerank s'accumulait sur un site, en étudiant le concept d'"énergie" du site (en fait quelque chose qui s'apparente à ce que les SEOs appellent dans leur jargon le "link juice" ou "jus de liens"). L'énergie d'un site est tout bonnement la somme de tous les pageranks des pages d'un site.

Ils ont conclu que le "link juice" présent sur un site (dans sa globalité) se calculait selon la formule très simple qui suit :

$$E_I = |I| + E_I^{in} - E_I^{out} - E_I^{dp}.$$

Où :

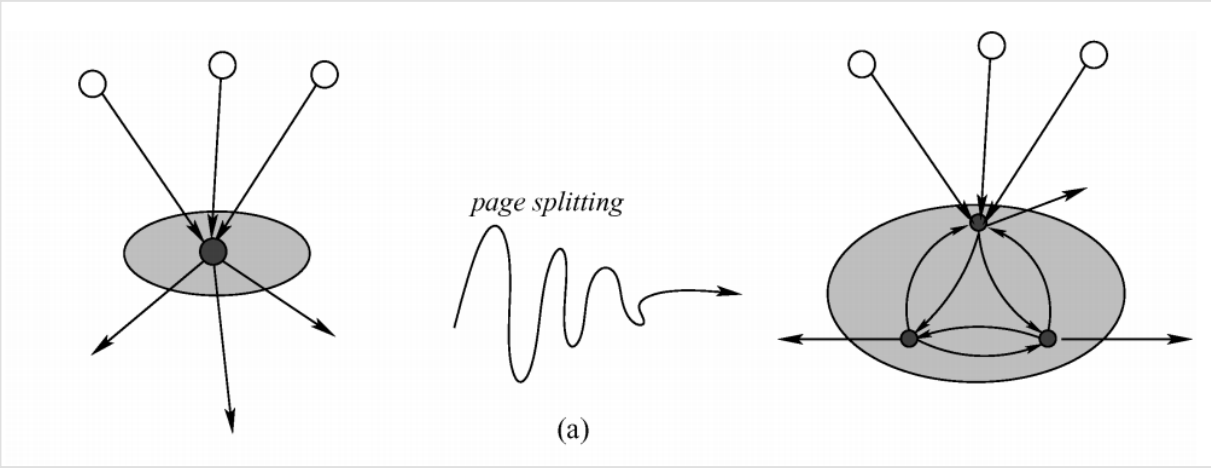
- E^{dp} correspond à l'énergie perdue dans les dangling pages ("dangling page" = une page qui ne comporte aucun lien sortant mais au moins un lien entrant : le "link juice" transmis à ces pages est perdu à 100%).

- E^{out} correspondant à la fuite de linkjuice produite par les liens sortants pointant vers d'autres domaines.

- E^{in} correspond quant à lui au linkjuice gagné grâce aux bakclinks qui transmettent du pagerank depuis des pages externes.

- Reste le facteur baptisé | I | qui représente en fait le pagerank "par défaut" produit par la transmission et l'accumulation du pagerank entre pages internes, grâce au maillage interne. Ce facteur | I | représente en fait... le nombre de pages du site.

Dans la pratique, un (très) gros site, démarrant avec des dizaines de millions de pages, peut théoriquement disposer d'un avantage et concentrer un pagerank non négligeable sur sa home page et ses principales pages de rubrique, là où un site de quelques pages partira avec un "linkjuice" quasi nul.



Un graphe tiré de l'article « Inside Pagerank ». Les chercheurs italiens démontrent que le fait de découper des pages en plusieurs pages plus petites conduit à une augmentation de l'énergie totale du site... Bref, en théorie, augmenter le nombre de pages augmente l'énergie (link juice) du site.

Précisons néanmoins que cet avantage fourni par le pagerank interne "par défaut" des gros sites disparait assez vite dès lors que les sites obtiennent un "link juice" important en provenance des backlinks, car le facteur | I | est souvent d'un ordre de grandeur rapidement très inférieur à E^{in} même pour les gros sites !

L'influence du maillage interne

Mais à "énergie" (linkjuice) égal, le maillage interne fournit deux avantages concurrentiels aux gros sites sur les petits :

- Leur surface sémantique est encore potentiellement augmentée par le maillage interne ;

- Localement, le maillage peut concentrer du link juice sur des pages importantes.

Une surface sémantique encore élargie par les anchor texts

Nous avons déjà signalé plus haut que disposer d'un grand nombre de pages permet à des sites de se positionner sur un plus grand nombre de requêtes qu'un petit site. Mais rappelons que les "anchor texts" ont un poids important dans l'algorithme. Sur un gros site, on va donc pouvoir disposer de dizaines, de centaines, voire de milliers de pages effectuant un lien interne vers une page donnée : faire "tourner" les ancres va donc permettre (en théorie) d'enrichir sémantiquement les pages sur de nouvelles expressions. Sur un petit site, le nombre de possibilités est considérablement plus limité. Il n'est pas rare de voir sur un site d'une quarantaine de pages beaucoup de pages profondes ne recevoir que deux ou trois liens.

Un maillage optimal peut donner un avantage aux pages des "gros sites"

Sur un gros site, on peut (théoriquement) "booster" certaines des pages en créant un maillage optimisé qui aura tendance à concentrer le linkjuice sur des pages importantes. Si on veut se positionner sur la requête "voiture d'occasion" avec un blog, on n'aura pas d'autres choix que d'aller à la chasse aux backlinks externes pour parvenir à se positionner. Mais pour un site d'annonces, les choses sont facilitées :
- Il dispose déjà de backlinks pointant probablement vers sa home page (c'est l'effet "gains de backlinks pour les gros sites" évoqué par Matt Cutts) ;
- Un maillage interne optimisé lui permet de diffuser du link juice depuis la home vers sa page rubrique "voiture d'occasion" ;

- Et, toujours en optimisant le maillage, on peut faire en sorte que toutes les annonces de la rubrique pointent vers la page de la rubrique parente, et participent à concentrer le pagerank sur la page correspondant à la rubrique "voiture d'occasion".

Le résulta est qu'une page d'un gros site qui reçoit zéro backlink peut, en théorie, être jugée plus importante qu'une page d'un petit site qui reçoit, lui, des backlinks, mais transmettant peu de link juice. Ce type de situation peut favoriser un gros site sur des requêtes moyennement concurrentielles où la plupart des pages qui répondent à la requête n'ont peu ou pas de backlinks (c'est à dire clairement des requêtes dites "longue traine").

Un biais de l'algorithme pour les gros sites ?

La conclusion que l'on peut tirer facilement en examinant tous ces mécanismes, c'est qu'il existe en théorie un biais dans l'algorithme, qui aura donc tendance à donner l'avantage aux pages de gros sites *versus* celles de petits sites, éventuellement au détriment de la pertinence.

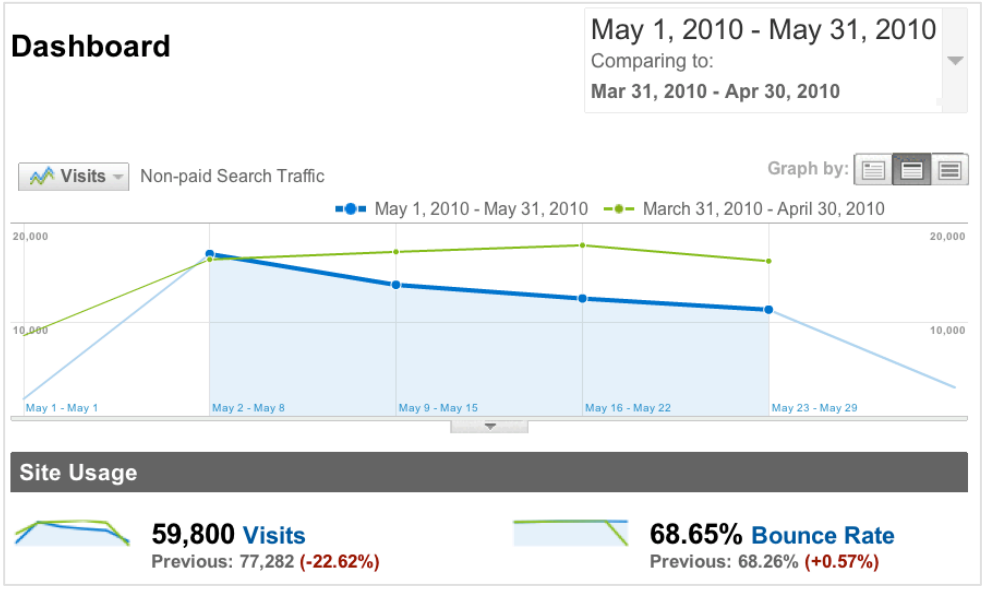
Pendant longtemps, ce biais théorique n'a pas réellement eu d'influence gênante sur la pertinence générale des résultats. Les raisonnements présentés ci-dessus n'impactent que certains des critères de classement, certes parmi les plus importants, mais il en existe bien d'autres. En outre, tant que la taille du site est représentative de la capacité d'un humain ou d'une entreprise à créer du contenu, cet avantage a eu plutôt tendance à favoriser des sites qui comptent vraiment sur le web.

Mais évidemment avec le temps et le développement du spam, la relation directe entre "importance réelle du site" et "taille du site" a eu tendance à devenir plus "floue". Et Google s'est rendu compte en particulier que sur les requêtes "longues traine", la domination des gros sites était un peu ... gênante.

Mayday : un correctif de l'algorithme sur la Longue Traîne, au détriment des gros sites

En mai 2010, Google a déployé un algorithme baptisé Mayday, qui a impacté le trafic longue traine d'un grand nombre de gros sites, parfois dans des proportions impressionnantes.

Mayday a été l'une des plus spectaculaires tentatives de Google pour corriger le biais en faveur des gros sites. Tous les observateurs ont noté que d'un seul coup, les pages internes de gros sites sans backlinks n'arrivaient plus à se positionner aussi bien sur des requêtes longue traine. Le "boost" accordé aux gros sites a donc fortement diminué. Mais dans la pratique (l'étude de SEO Chat le prouve) avoir un site avec beaucoup de pages représente toujours un avantage concurrentiel.



L'impact typique de Mayday sur un « gros » site : la perte du trafic sur les requêtes « long tail » a généralement conduit à des pertes importantes de trafic, parfois de l'ordre de 20 à 30% du trafic organique.

Dois-je augmenter la taille de mon site pour émaliorer mon référencement ?

Evidemment, on pourrait en conclure qu'augmenter le nombre de pages de son site est la solution ultime pour doper son référencement.

Ce qui est clair, c'est qu'augmenter la "surface sémantique" de votre site en créant des pages constituant de nouvelles "bonnes" pages d'atterrissage sur de nouvelles requêtes est souvent une bonne idée. Mais n'oubliez pas que doubler le nombre de pages ne veut pas forcément dire doubler votre trafic SEO : le trafic n'augmentera que si vos nouvelles pages sont 1) indexées, et 2) correctement positionnées.

Dans la pratique, c'est parfois très compliqué : ce qui est difficile, c'est de créer un grand nombre de pages de qualité, pertinentes, qui rendent véritablement service à l'internaute. Rappelons que les "fermes de contenus" (dont l'archétype était le site eHow) avaient compris qu'en créant autant d'articles que de requêtes populaires, on pouvait exploiter ce biais de l'algorithme pour les "gros sites" et squatter les premières pages sur des requêtes longue traine. Sauf que Google a trouvé la parade avec le filtre Panda, et que ce modèle est devenu ... dangereux.

En ce qui concerne l'exploitation du maillage interne, le "boost" ne s'obtient que si le site contient un très grand nombre de pages (des dizaines de milliers). Or, lorsque l'on est confronté à des sites "vitrines", ou des sites "corporate", l'idée de les transformer en sites comportant des milliers de pages peut laisser perplexe. La solution classique consiste à créer un "blog", mais le nombre de pages que l'on peut créer ainsi reste limité, et les efforts à accomplir pour publier des centaines de billets pertinents sont à peser face à l'impact beaucoup plus direct et positif d'une bonne campagne de promotion (pour obtenir des "leads" et des backlinks).

En réalité, la taille d'un site est le plus souvent un état de fait pour le référenceur qu'une "variable d'ajustement". C'est une question d'ordre de grandeur : on peut certes doubler la taille d'un site, mais faire passer un site vitrine de 10 pages à un site de 10 millions de pages est rarement une stratégie réaliste.

Qui plus est, on peut d'ores et déjà se demander si cette stratégie continuera d'être payante avec la montée des requêtes en langage naturel et l'indexation des concepts et des entités. Google prétend qu'Hummingbird, son nouvel algorithme dévoilé en septembre dernier, est censé favoriser les sites de niche. Ce n'est pas vraiment flagrant aujourd'hui. Mais qu'en sera-t'il demain ? Ce qui est clair, c'est qu'on ne peut plus espérer se positionner durablement avec un site artificiellement "gros"... Mais dans la jungle que constitue le web, la loi du plus fort, du plus gros, et du plus puissant existe bel et bien, et les classements de Google reflètent cet état de fait, en toute logique...

Bibliographie

Le "thread" de WebmasterWorld :
Has Google increased the importance of website size as a ranking factor ?
http://www.webmasterworld.com/google/4591155.htm

L'étude de SEO Chat de 2010
http://www.seochat.com/c/a/choosing-keywords-help/keyword-difficulty-vs-size-of-domain/

L'étude de websitemagazine
http://www.websitemagazine.com/content/blogs/posts/archive/2010/02/09/seo-research-website-size-study.aspx

Vidéos de Matt Cutts

Does a site rank better if it has a lot of indexed pages? (28 octobre 2013)
http://www.youtube.com/watch?v=AVOrml7fp2c

Does the size of a website affect its authority in Google? (21 octobre 2009)
http://www.youtube.com/watch?v=Mha9q2aAfdM

Publications scientifiques

Inside Pagerank
(Monica Bianchini, Marco Gori et Franco Scarselli), Université de Sienne
http://dl.acm.org/citation.cfm?id=1052938
http://www.cs.bham.ac.uk/~pxt/IDA/pagerank.pdf

Philippe YONNET , ***Directeur de l'agence Search-Foresight / Groupe MyMedia. Président de l'association SEO Camp (http://www.seo-camp.org/)***