The Connectivity Server: fast access to linkage information on the web

Bharat, Broder, Henzinger,Kumar, Venkatasubramaniam

Avril 1998

Introduction

- Connaître les informations sur les links entrants et sortants depuis une page web permet de mieux exploiter les données.
- Jusqu'à 1998, cette collecte d'information n'avait été faite que dans un voisinage restreint de certains sites et en utilisant une technique lourde et peu efficace.
- Idée: créer un serveur de connectivité à partir des pages référencées par un moteur de recherche (Altavista).

But du serveur de connectivité

- Mettre en évidence les résultats énoncés par Kleinberg (soit l ’existence pour un domaine donné de sites « Authorités » et de sites « Hubs »)
- Améliorer les moteurs de recherche en proposant à l ’utilisateur lambda les sites cité plus haut comme base de recherche.
- Donner une représentation des liens et des nœuds d’une partie importante du WWW

Le serveur de connectivité (1)

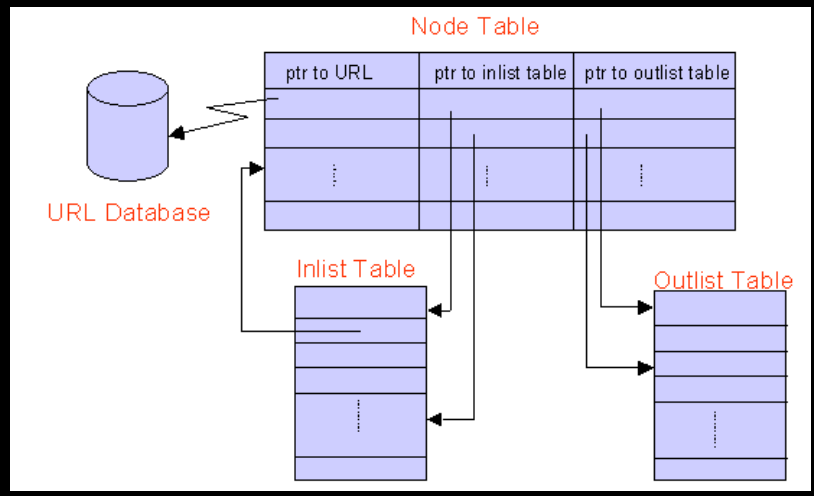
- Utilise la base de donnée de Alatavista
- Permet de rechercher TOUS les liens entrants et sortants d'un point du web quelconque.
- Produit le voisinage entier jusqu'à une profondeur n quelconque.
- Algorithme de recherche efficace.

Le serveur de connectivité (2)

- Représenter un graphique de quelques nœuds est trivial, un graphique de quelques 100 millions de nœuds et près de 1 milliard de liens un défit technologique.
- Il s'agit de bien penser la structure de données ainsi que son accès.
- Cette structure doit intégrer l'aspect « évolutionniste » des éléments du WWW

Fonctionnement (1)

- Chaque nœud est un pointeur sur la base Altavista
- Chaque nœud (URL) possède une liste de liens entrants et sortants (pointeurs sur la même table).
- Si « bêtement »: 8 Go

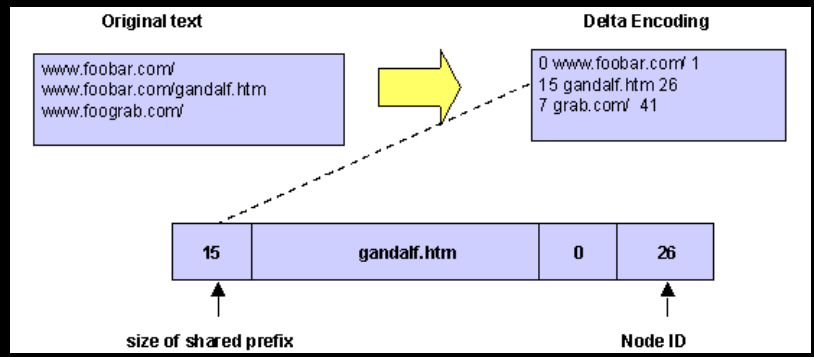


Fonctionnement (2)

- Chaque nœud représente une URL
- La méthode de pointeurs d ’URL fait gagner 70 % de la place disque, mais ralenti la recherche (effectuer la recherche comlpète en suivant les pointeurs)
- Optimisation: utilisation d’URL’s « balises » (points de contrôle) complètes au lieu de pointeurs.

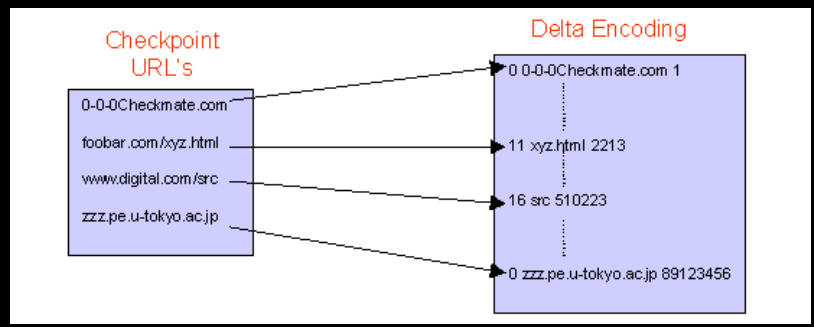
Fonctionnement (3)

- Utilisation de « delta » par rapport à une page pointée et à la page courante. C'est un gain de place.
- Traduction: delta - adresse complète.



Fonctionnement (4)

- Utilisation des points de contrôle.



Fonctionnement (5)

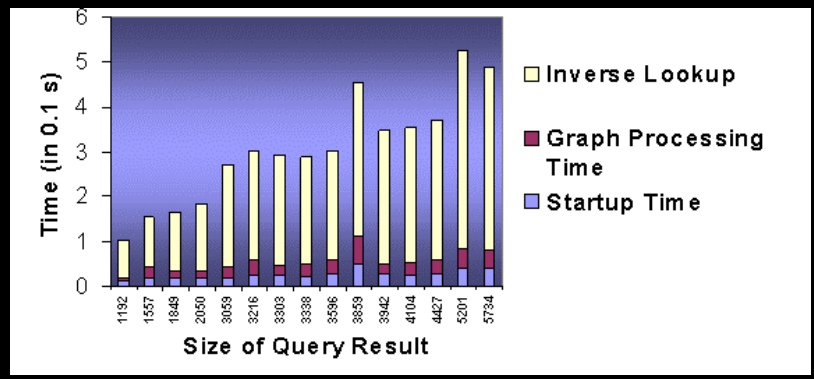
- La structure de donnée est pensée pour que l ’on puisse ajouter ou supprimer des nœuds
- Pour ce faire, la structure de données est plus grande que celle qui permettrait dans un premier temps à stocker au temps t=0 la base de données de AltaVista.
- Mise à jour: tous les jours, une fois la structure temporaire crée, on reconstruit la table.

Exécution (1)

- Le serveur procède en trois étapes pour traiter les demandes:
- TRADUIT URL - URL id
- PARCOURS la structure pointée
- TRADUIT URL id - URL
- Le temps de calcul (DEC Alpha 300 MHz, 4 G0 Ram) prend environ 0,1 ms / URL
- Temps de parcours: 0,01 ms / URL

Exécution (2)

- Voisinages de 1100 à 5800 URL (liens)
- 80 % du temps de calcul par requête est utilisé pour la traduction « inverse » (soit URL id - URL)



Applications réelles (1)

- Le serveur de connectivité a été implémenté en 1998 (pas trouvé de lien actuel) et possédait un lien sur le site principal de AltaVista.

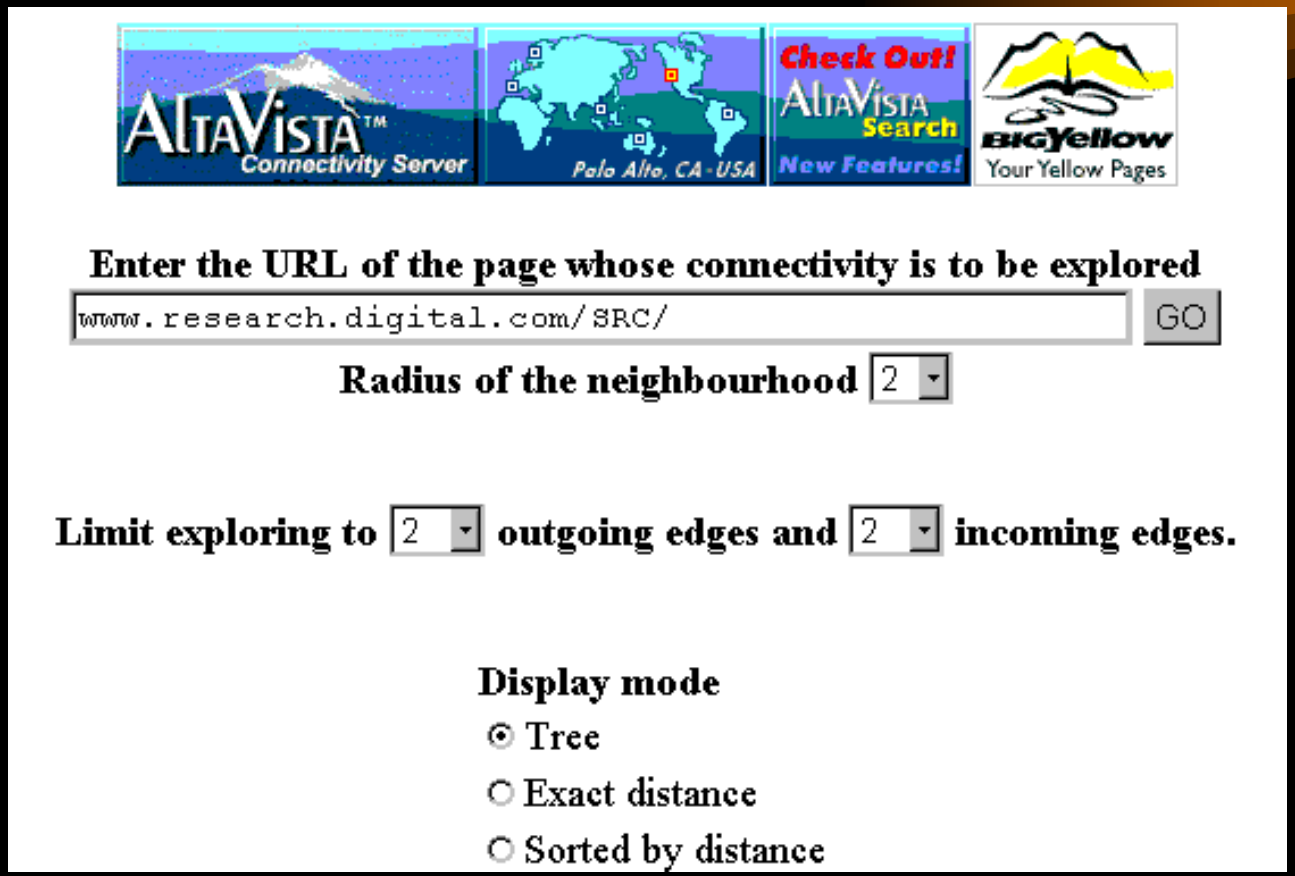
Applications réelles (2)

- Exemple de requête: site www.digital.com
- Sélection des liens entrants et sortants dans l ’interface de base (non étendue)

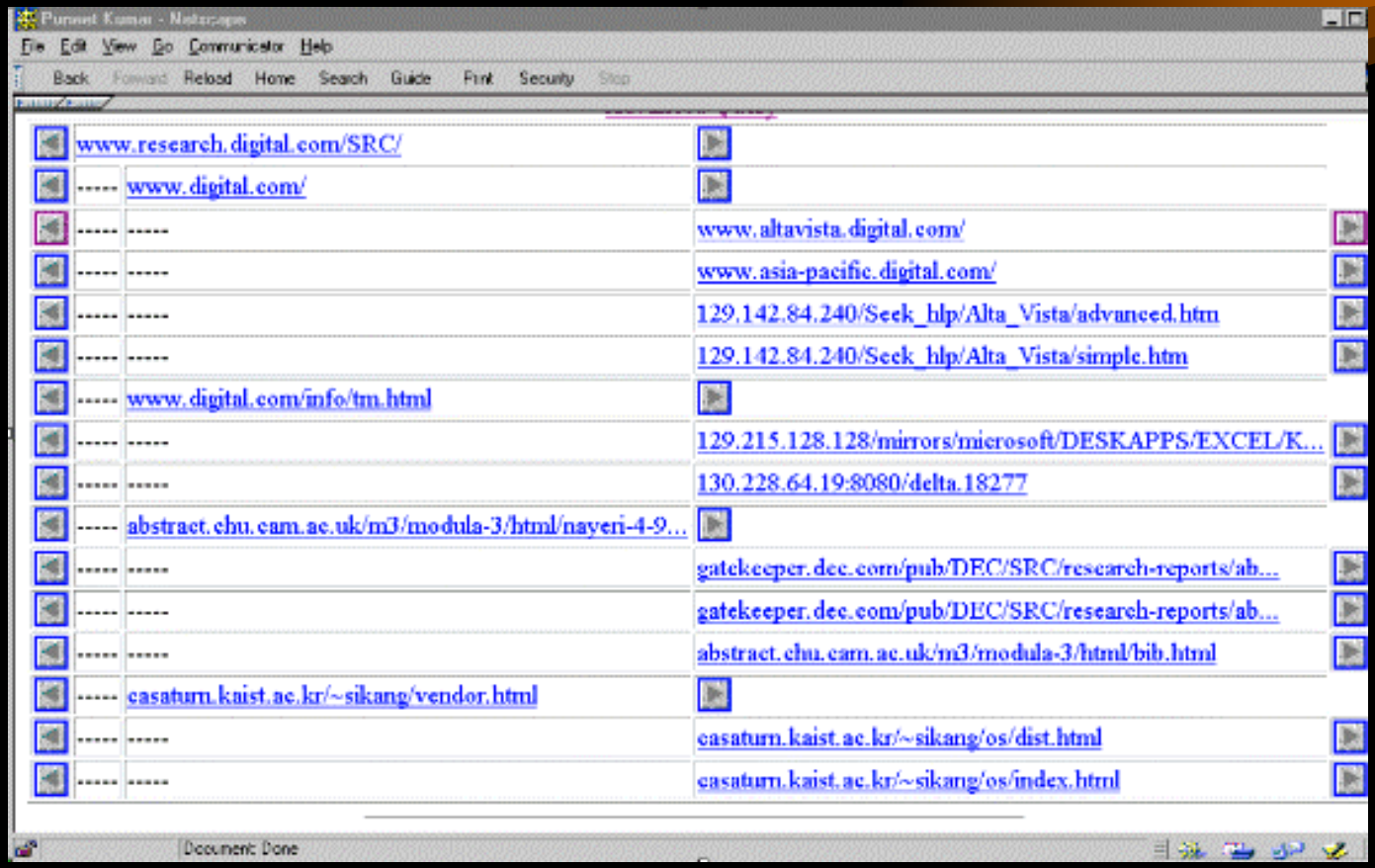
Applications réelles (3)

- Interface de requête avancée.
- Profondeur de recherche (profondeur des liens depuis une page donnée)
- Limite du nombre de lien entrants ou sortants
- Représentation sous forme d'arbre, de graphique ou de liste du résultat de la requête

Applications réelles (4)



Applications réelles (5)

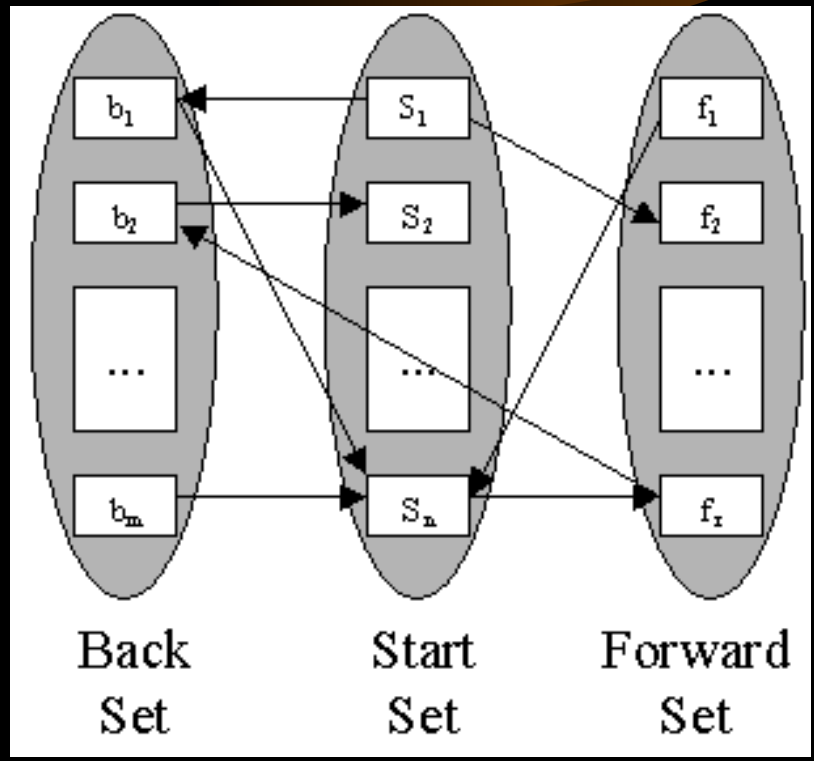


Affichage de la connectivité (1)

- Autre application plus complexe.
- Le serveur de connectivité calcule l ’intégralité du voisinage d ’un nœud WWW (au sens mathématique)
- Affichage de la connectivité signifie le calcul du voisinage à 1 pour les liens entrants et sortants.

Affichage de la connectivité (2)

- Start Set = pages de base
- Back Set = pages qui possèdent un lien sur l ’une au moins des pages du Start Set
- Forward Set = pages pointées par une page du Start Set

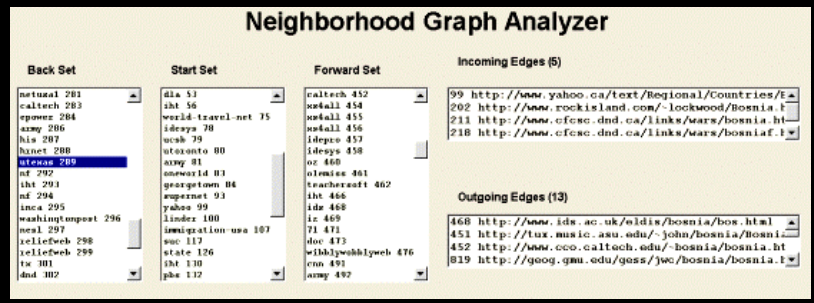


Affichage de la connectivité (3)

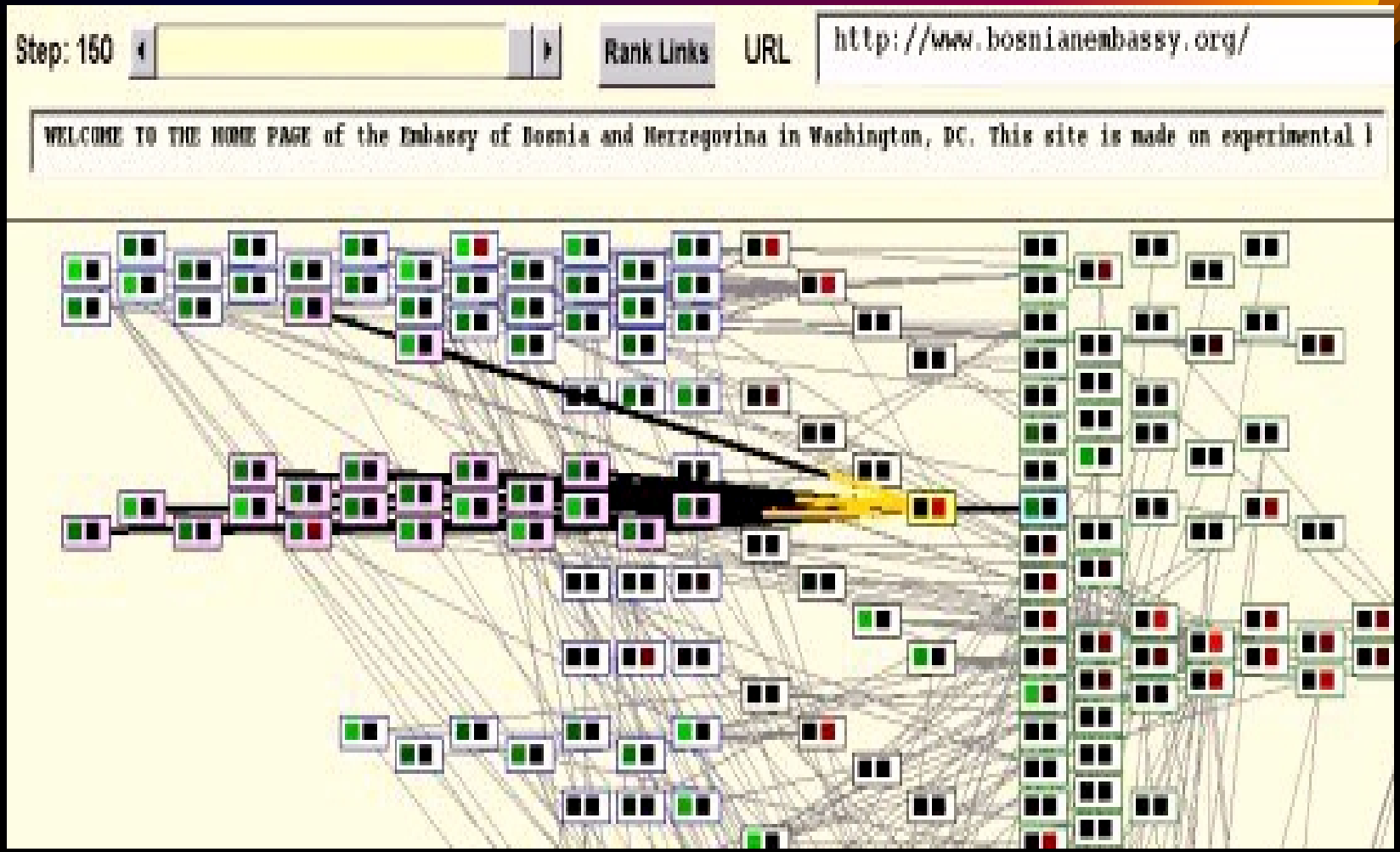
- Le calcul se fait grâce à la structure de données.
- Le temps de calcul prend quelques minutes. Il s ’agit dans un premier temps de parcourir la structure de données, puis de filtrer les liens qui intéressent. Soit ceux qui sont en relation avec la requête

Affichage de la connectivité (4)

- Affiche les liens entrants et sortant
- A partir d ’un ensemble de pages données
- Affiche textuellement le voisinage.
- Affichage graphique

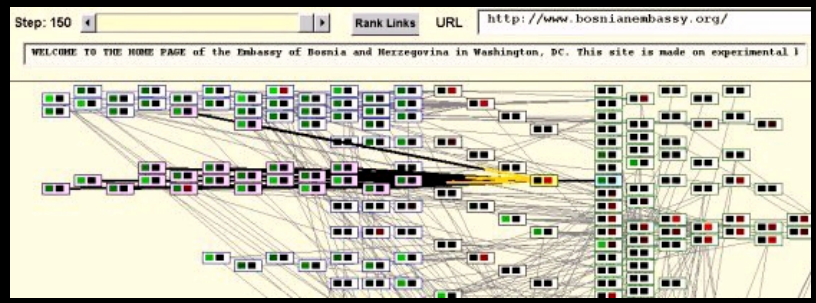


Affichage de la connectivité (5)



Affichage de la connectivité (6)

- Affiche deux « boîtes » l ’une rouge et l ’autre verte
- Rouge = authorités
- Vert = hubs
- Un nœud sélectionné permet d ’afficher la direction des liens



Analyse de la connectivité (1)

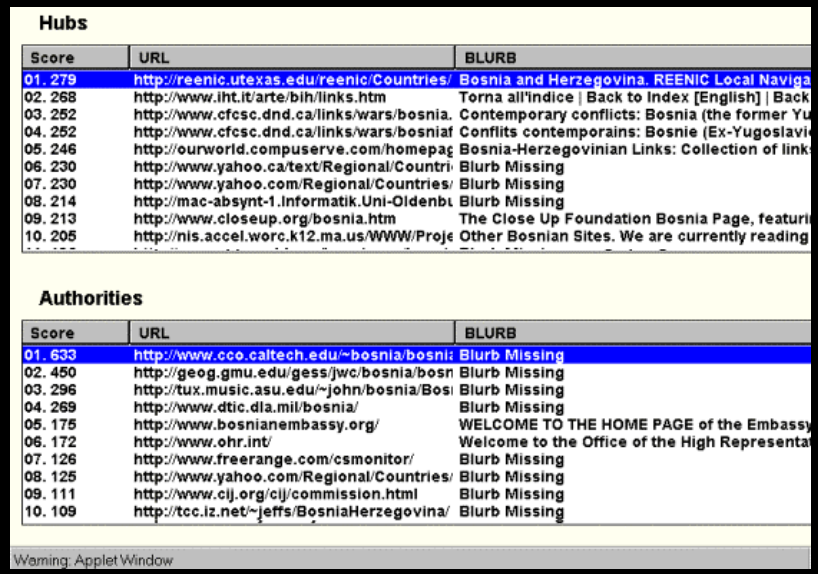
- Le but est donc atteint: recherche des pages utiles comme « autorités » et celles utiles comme « hubs ».
- Chaque page a un ranking comme autorité (A) et comme hub (H).
- Le parcours du voisinage permet de modifier ces valeurs (plus une page est pointée par celles du voisinage, plus son ranking A est élevé, etc..)

Analyse de la connectivité (2)

- Dans la pratique, les tests du serveur de connectivité ont été faits de la manière suivante:
- On prend les 200 meilleurs résultats d ’une requête d ’utilisateur sur AltaVista comme Start Set
- Cela permet à partir d ’une requête donnée de donner les Hubs et Autorité pour cette requête.

Analyse de la connectivité (3)

- Résultat: on peut donner pour la requête « Bosnie » les pages suivantes (classées selon leur ranking Autorité et Hub (A et H)



Hubs

Score	URL	BLURB
01. 279	http://reenic.utexas.edu/reenic/Countries/	Bosnia and Herzegovina. REENIC Local Naviga
02. 268	http://www.iht.it/arte/bih/links.htm	Torna all'indice | Back to Index [English] | Back
03. 252	http://www.cfcsc.dnd.ca/links/wars/bosnia.	Contemporary conflicts: Bosnia (the former Yu
04. 252	http://www.cfcsc.dnd.ca/links/wars/bosniaf	Conflits contemporains: Bosnie (Ex-Yougoslavi
05. 246	http://ourworld.compuserve.com/homepag	Bosnia-Herzegovinian Links: Collection of link
06. 230	http://www.yahoo.ca/text/Regional/Countri	Blurb Missing
07. 230	http://www.yahoo.com/Regional/Countries/	Blurb Missing
08. 214	http://mac-absynt-1.informatik.Uni-Oldenbu	Blurb Missing
09. 213	http://www.closeup.org/bosnia.htm	The Close Up Foundation Bosnia Page, featuri
10. 205	http://nis.accel.worc.k12.ma.us/WWW/Proje	Other Bosnian Sites. We are currently reading

Authorities

Score	URL	BLURB
01. 633	http://www.cco.caltech.edu/~bosnia/bosni	Blurb Missing
02. 450	http://geog.gmu.edu/gess/jwc/bosnia/bosn	Blurb Missing
03. 296	http://tux.music.asu.edu/~john/bosnia/Bosi	Blurb Missing
04. 269	http://www.dtic.dla.mil/bosnia/	Blurb Missing
05. 175	http://www.bosnianembassy.org/	WELCOME TO THE HOME PAGE of the Embassy
06. 172	http://www.ohr.int/	Welcome to the Office of the High Representa
07. 126	http://www.freerange.com/csmonitor/	Blurb Missing
08. 125	http://www.yahoo.com/Regional/Countries/	Blurb Missing
09. 111	http://www.cij.org/cij/commission.html	Blurb Missing
10. 109	http://tcc.iz.net/~jeffs/BosniaHerzegovina/	Blurb Missing

Warning: Applet Window

Conclusion

- Le serveur de connectivité met en évidence les résultats de Kleinberg.
- Il permet de rechercher efficacement les Autorités et les Hubs pour une requête donnée
- Il permet donc de simplifier la recherche d ’information pour l ’utilisateur lambda
- Il est utile à des fins de marketing