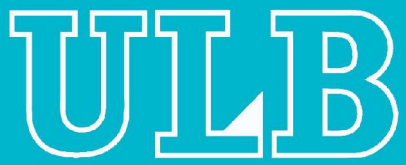
Internet et les systèmes d'informations intégrés

Moteurs de Recherche



Contenu

- Introduction
- Représentation
- Traitement
- Recherche
- Améliorations
- Méta-moteurs
- Référencement
- Conclusions

Contenu

- Introduction
- Représentation
- Traitement
- Recherche
- Améliorations
- Méta-moteurs
- Référencement
- Conclusions

Introduction

- Internet est la principale source d'information pour beaucoup :
  - Moteurs sont les outils les plus utilisés pour rechercher de l'information.
  - Introduction technique.
  - Peu d'informations sur les algorithmes réellement utilisés par les moteurs de recherche.
- Internet est un hyper-espace :
  - Les documents sont inter-connectés.
  - Comment exploiter ces liens?

Contenu

- Introduction
- Représentation
- Traitement
- Recherche
- Améliorations
- Méta-moteurs
- Référencement
- Conclusions

Représentation (1/3) - Pourquoi

- Logiciels (ex: Moteurs de recherche) nécessitent une représentation des documents.
- Logiciels n'ayant pas de capacité cognitive, il faut introduire des modèles :
  - Représentation synthétique.
  - Simplification inévitable.
- Plusieurs modèles existent.
- Actuellement, la plupart des représentations ne tiennent compte que des informations textuelles.

Représentation (2/3) - Mots-clés (1/2)

- Méthode la plus utilisée: Documents sont représentés par des mots-clés.
- Supposer l'indépendance des mots (ex: Présence du mot "Deep" ne dit rien sur la présence du mot "Purple") :
  - Peut sembler une limitation.
  - Modèles existants ne donnent pas de meilleur résultats.
- Associer des poids à des mots-clés pour définir une importance.

Représentation (3/3) - Mots-clés (2/2)

- Besoin :
  - Identifier les mots-clés → Traitement.
  - Calculer les poids → Modèle.
- Exemple d'un document parlant des Beatles :

Mots	Poids
McCartney	5
Lennon	5
Harrisson	4
Ringo	1
help	2
rock	3

Contenu

- Introduction
- Représentation
- Traitement
- Recherche
- Améliorations
- Méta-moteurs
- Référencement
- Conclusions

Traitement (1/14) - Contenu

- Pourquoi?
- Analyse Lexicale
- Stop-liste
- Désuffixation
- Thésaurus
- Exemple

Traitement (2/14) - Pourquoi? (1/2)

- But du traitement est de sélectionner les mots-clés.
- Objectif sémantique: tous les mots ne sont pas nécessaires.
- Objectif informatique est de limiter le nombre de mots-clés :
  - Limiter la quantité de mémoire (vive et disques durs).
  - Limiter le temps de traitement.

Traitement (3/14) - Pourquoi? (2/2)

- Traitement peut-être combinaison :
  - Analyse Lexicale.
  - Stop-liste.
  - Désuffixation.
  - Thésaurus
- Certains moteurs de recherche n'utilisent aucun traitement et indexent donc tous les "mots".

Traitement (4/14) - Analyse Lexicale (1/5)

- Transformer une séquence de caractères en "mots" en utilisant une liste de règles :
  - Temps de calcul augmente avec nombre de règles.
  - Certains moteurs acceptent toutes les séquences.
- 4 classes de séquences de caractères :
  - Nombres.
  - Traits d'union.
  - Ponctuations.
  - Casse.

Nombres

- Les nombres seuls n'ont pas beaucoup d'intérêt, on peut donc ne pas les utiliser.
- Certains mots contenant des nombres ont une importance (ex: ISO9002, Z39.50, ...).
- Solution simple: retenir que les mots commençant par une lettre (mais: 510B.C).
- Généralement établir une série de règles identifiant les mots contenant des nombres et qui seront utilisés.

Traits d'union

- Lorsque traits d'union rencontrés, il faut faire un choix.
- Enlever et considérer "state-of-the-art" comme identique à "state of the art".
- Problèmes :
  - Certains mots sont différents associés que séparés (ex: "porte-manteau").
  - Utiliser pour gérer les césures.
- Utiliser une série de règles.

Traitement (7/14) - Analyse Lexicale (4/5)

Ponctuations

- Séparation de phrases.
- Mais, utiliser parfois dans un mot (ex: TCP/IP, X25, pfrancq@ulb.ac.be, ...).
- Enlever et considère "pfrancq@ulb.ac.be" comme "pfrancqulb.ac.be".
- Utiliser une série de règles.

Traitement (8/14) - Analyse Lexicale (5/5)

Casse

- La casse des mots a souvent peu d'importance.
- Transformer tous les mots en minuscules et majuscules.
- Problème : Certains mots ont une signification différente avec différentes casses (ex: "Bank" et "bank").
- Utiliser une série de règles.

Traitement (9/14) - Stop-liste (1/2)

- Certains mots d'une langue apparaissent très souvent.
- Stop-liste : liste de mots apparaissant très souvent mais n'ayant pas beaucoup de contenu sémantique (ex: le, la, en, de, ...).
- Études pour l'anglais et le français montrent que 20 à 30% des mots dans les documents sont de la stop-liste.

Traitement (10/14) - Stop-liste (2/2)

- Ne pas utiliser comme mots-clés les mots de la stop-liste.
- Stop-liste peut également contenir des verbes, des adverbes ou encore des adjectifs.
- Problème : Peut détruire de l'information (ex: "to be or not to be").

Traitement (11/14) - Désuffixation

- Certains mots différents proviennent des mêmes radicaux.
- Considérer tous les mots d'un même radical comme le même mot-clé.
- Algorithme de désuffixation :
  - Utiliser un ensemble de règles pour transformer un mot en racine.
  - Garder que les racines:

ex: chienne ⇒ chienn ⇒ chien

Traitement (12/14) - Thésaurus

- Thésaurus est un ensemble de mots ayant une grande valeur discriminatoire
- Utiliser comme mots-clés uniquement les mots repris dans le thésaurus
- Aider l'utilisateur à choisir les bons mots-clés pour la recherche.
- Problème : Thésaurus est orienté domaine.

Traitement (13/14) - Exemple (1/2)

1

Supplément Multimédia Vendredi (31 décembre1999)

MP3, comment ça marche?

Pour lire les fichiers MP3, il est indispensable de se procurer un petit logiciel. Il en existe d'innombrables versions sur le Net. Le plus connu est certainement Winamp, que l'on peut se procurer à l'adresse http://www.winamp.com. Pour emporter ses sélections musicales, il est possible de les transférer sur un petit baladeur comme le Diamond Rio (www.diamondmm.com) ou l'un de ses nombreux concurrents. Le site Web le plus célèbres consacrés à la distribution de ces fichiers est www.mp3.com qui permet de télécharger des fichiers libres de droit pour découvrir l'univers du MP3. Enfin, pour trouver une chanson ou un artiste parmi plus d'un million de fichiers musicaux, le moteur de recherche de Lycos offre un outil spécialisé (http://mp3.lycos.com). ...

2

Supplément Multimédia Vendredi décembre

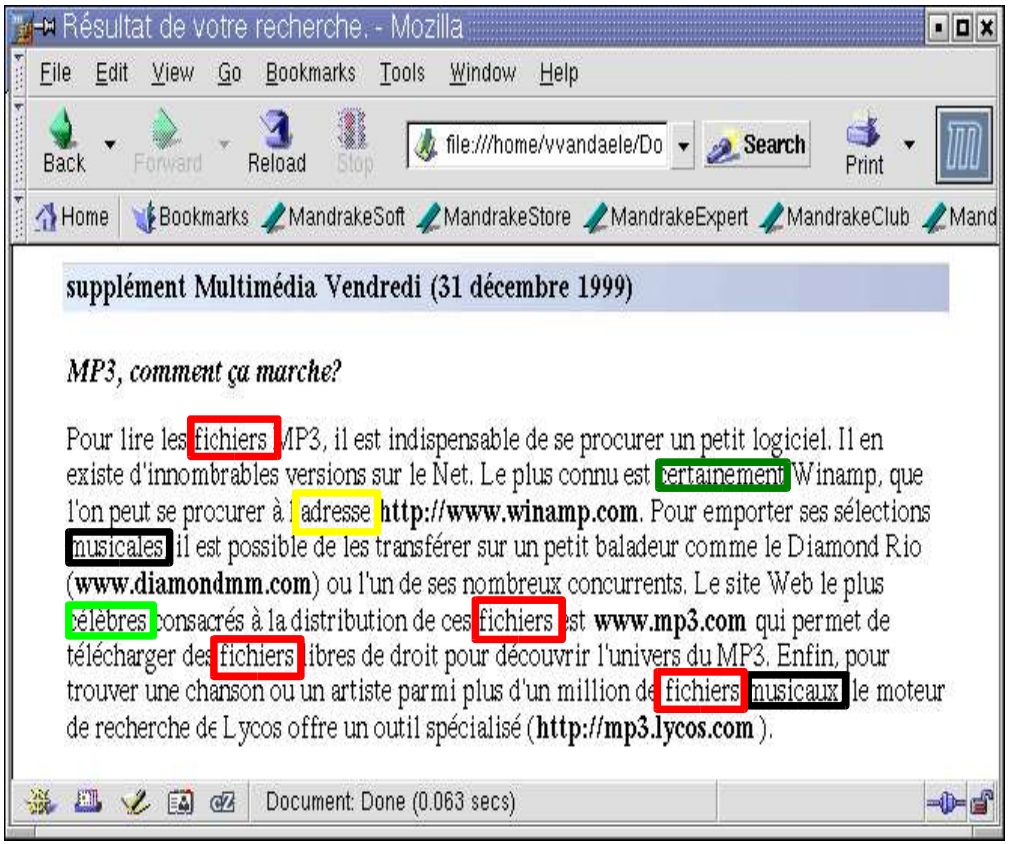
marche

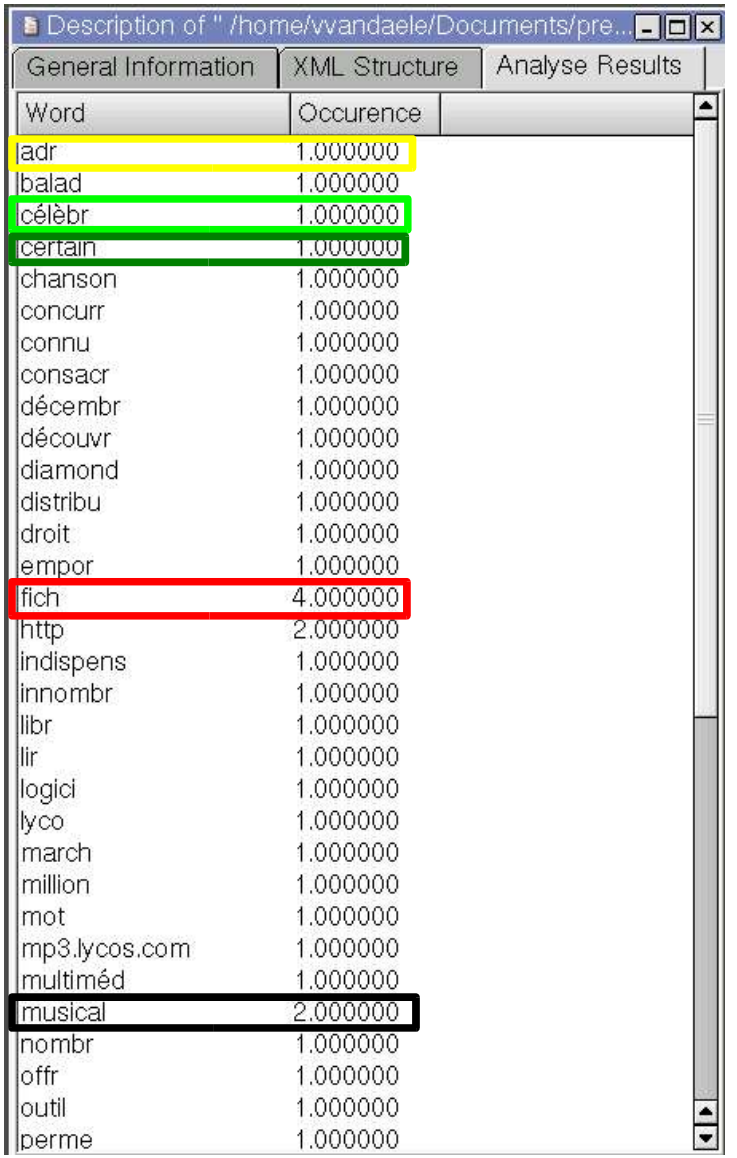
lire fichiers indispensable procurer petit logiciel. existe innombrables versions Net. connu certainement Winamp, peut procurer adresse http://www.winamp.com. emporter sélections musicales, possible transférer petit baladeur comme Diamond Rio (www.diamondmm.com) nombreux concurrents. site Web célèbres consacrés distribution fichiers www.mp3.com permet télécharger fichiers libres droit découvrir univers. trouver chanson artiste million fichiers musicaux, moteur recherche Lycos offre outil spécialisé (http://mp3.lycos.com). ...

3

Appliquer la désuffixation

Traitement (14/14) - Exemple (2/2)





Contenu

- Introduction
- Représentation
- Traitement
- Recherche
- Améliorations
- Méta-moteurs
- Référencement
- Conclusions

Recherche (1/6) - Contenu

- Modèle Booléen
- Modèle Vectoriel

Recherche (2/6) - Modèle Booléen (1/2)

- Poids des mots :
  - 0 si pas présent dans le document.
  - 1 si présent.
- Utilisateur entre une requête formée par des mots et des opérateurs booléens.

  'deep' and 'purple'
- Moteurs de recherche renvoient tous les documents répondant à la requête.

Recherche (3/6) - Modèle Booléen (2/2)

- Requête basée sur mots 'beatles' et 'rock' :
  - 'beatles' and 'rock' → 1 seul document "Beatles" trouvé.
  - 'beatles' or 'rock' → Tous les documents trouvés.

	Beatles	Beatles2	Deep Purple
beatles	1	1	0
deep	0	0	1
guitar	1	1	1
lennon	1	0	0
mccartney	1	1	0
purple	0	0	1
rock	1	0	1

Recherche (4/6) - Modèle Vectoriel (1/3)

- Poids des mots peut prendre n'importe quelle valeur numérique.
- Poids représente la valeur discriminatoire d'un mot :
  - Mots apparaissant dans tous les documents n'ont pas de valeur discriminatoire.
  - Mots apparaissant dans un seul document n'ont pas de valeur discriminatoire.
  - Mots intéressants sont ceux qui apparaissent très souvent dans un sous-ensemble de documents.

Recherche (5/6) - Modèle Vectoriel (2/3)

- Requête est composée de deux mots 'beatles rock'.
- Similarité entre documents et requête :
  - sim(Beatles,requête)=0.8
  - sim(Beatles2,requête)=0.6
  - sim(Deep Purple,requête)=0.2

Mots	Beatles	Beatles 2	Deep Purple
beatles	3,1	2,2	0
deep	0	0	2,3
guitar	0,7	0,6	0,8
lennon	1,2	0	0
mccartney	2,13	2,6	0
purple	0	0	1,9
rock	0,8	0,5	0,6

Recherche (6/6) - Modèle Vectoriel (3/3)

- Fixer un seuil maximale de similarité pour considérer qu'un document est intéressant (ex: 0.5).
- Classer les documents par similarité décroissante :
  1. Beatles.
  2. Beatles2.

Contenu

- Introduction
- Représentation
- Traitement
- Recherche
- Améliorations
- Méta-moteurs
- Référencement
- Conclusions

Améliorations (1/11) - Contenu

- Autorités
- Méta-données
- Concepts
- Feedback
- Liens

Améliorations (2/11) - Autorités

- Utiliser des "autorités morales" pour mieux déterminer l'intérêt d'un document (ex: Google).
- Certains sites sont plus sérieux et peuvent "valider" un document :
  - Quand l'information se trouve chez eux, elle est plus intéressante qu'ailleurs.
  - Quand ils réfèrent d'autres sites, ceux-ci devraient avoir de l'information intéressante.
- Faire intervenir ces autorités dans le classement en proposant d'abord les documents validés.

Améliorations (3/11) - Méta-données

- Certains documents contiennent des méta-données, c'est-à-dire des données sur les données (ex: Date de création, auteurs, ...).
- Les informations contenues dans les méta-données sont très importantes.
- Considérer qu'il faut chercher principalement dans les méta-données.

Améliorations (4/11) - Concepts

- La plupart des modèles supposent l'indépendance des mots.
- Existe des modèles qui permettent de construire des concepts (ensemble de mots) :
  - Étudier les co-occurrences des mots (ex: Poids non nul pour des mots qui n'apparaissent pas).
  - Utiliser les distances lexicales (ex: Mots 'deep' et 'purple' sont souvent ensemble → Concept 'deep purple').

Améliorations (5/11) - Feedback (1/2)

- Lorsque l'utilisateur recherche de l'information, il n'a pas une connaissance parfaite de celle-ci :
  - Les mots-clés utilisés ne sont pas forcément adaptés.
  - Certains mots-clés pouvant être intéressants ne sont pas utilisés.
- Tous les documents renvoyés par le moteur de recherche ne sont pas pertinents (même si un certain nombre le sont).

Améliorations (6/11) - Feedback (2/2)

- User Relevance Feedback permet d'améliorer la pertinence des documents proposés.
- Méthode :
  1. Utilisateur utilise une première requête.
  2. Le moteur lui renvoie une série de documents.
  3. Utilisateur coche parmi les 20 ou 30 mieux classés, ceux qui sont réellement intéressants.
  4. Le système construit automatiquement une nouvelle requête et renvoie une nouvelle série de documents.
  5. Retour à l'étape 3.

Contexte (1/2)

- Utiliser structure en hyperliens.
- Hyperliens représentent des jugements implicites humains :
  - Lorsqu'un créateur de page *p* inclus un lien vers *q*, il confère une certaine autorité à *q* :
    - Lien sortant de p.
    - Lien entrant de q.
  - Lorsqu'une page est souvent considérée comme autorité, elle est sans doute intéressant.

Contexte (2/2)

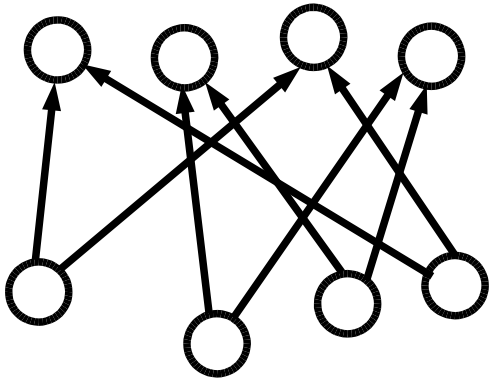
- Problème : Souvent des liens ne sont utilisés qu'à des fins de navigation (ex: Home page).
- Deux types de liens :
  - Intrinsèques : Liens pointant vers le même domaine.
  - Transversaux : Liens pointant vers d'autres domaines.
- Solution : Ne tenir compte que des liens transversaux.

Méthode Basique

- Prendre toutes les pages satisfaisant une requête et ayant un grand nombre de liens entrants.
- Problèmes :
  - Les pages les plus intéressantes ne répondent pas forcément à la requête (ex: Sites automobiles ne contiennent pas les mots 'constructeurs automobiles').
  - Des pages génériques (ex: Google) seraient des autorités pour n'importe quelle requête.

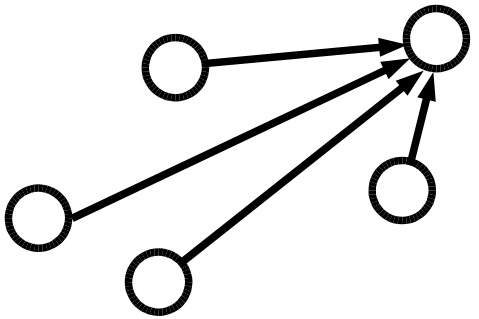
Autorités et Hubs

- On peut distinguer deux types de pages:
  - Autorités : Pages ayant des liens venant de hubs.
  - Hubs : Pages ayant des liens vers des autorités.
- Relations entre les autorités et les hubs renforcent leur rôle respectif.



Autorités

Hubs



Méthode

- Algorithmes de détermination des hubs et des autorités dérivent de la théorie des graphes.
- Méthode :
  1. Utilisateur entre une requête.
  2. Un moteur de recherche est utilisé pour renvoyer une première série de documents.
  3. Les documents les mieux classés seront utilisés.
  4. Déterminer les hubs et les autorités.
  5. Proposer les autorités à l'utilisateur.

Contenu

- Introduction
- Représentation
- Traitement
- Recherche
- Améliorations
- Méta-moteurs
- Référencement
- Conclusions

Méta-moteurs (1/6) - Contenu

- Principe
- Méthode
- Choix des moteurs
- Classement
- Problèmes

Méta-moteurs (2/6) - Principe

- Chaque moteur est «unique» :
  - Algorithmes de recherche propres (et propriétaires).
  - Index partiel des documents Internet.
  - Classement propre des documents trouvés.
- Estimations récentes indiquent un taux de recouvrement de 60%.
- Combiner recherches obtenues par plusieurs moteurs.

Méta-moteurs (3/6) - Méthode

- Principe de la méthode :
  1. L'utilisateur entre une requête.
  2. Lancer la requête sur plusieurs moteurs.
  3. Consolider les résultats en classer les documents.
- Certains méta-moteurs permettent :
  - Définir des requêtes à lancer régulièrement.
  - Faire une recherche incrémentale.
- Problèmes :
  - Quels moteurs choisir?
  - Comment classer?

Méta-Moteurs (4/6) - Choix des moteurs

- Deux approches pour le choix :
  - Tous les moteurs connus.
  - Les moteurs que «l'on sait les plus pertinents».
- Inconvénients :
  - Temps et coût proportionnels aux nombre de moteurs.
  - Pertinence est souvent basée sur l'existence de méta-données concernant les moteurs.
- Piste de recherche :
  - Évaluer la pertinence pour chaque requête.
  - Utiliser un échantillon de réponse (2-3 documents).

Méta-moteurs (5/6) - Classement

- Classer la liste de tous les documents trouvés par les moteurs.
- Plusieurs méthodes pour classer les documents :
  - Recalculer «localement» une similarité.
  - Mettre les documents apparaissant le plus souvent dans les réponses en tête de liste.
  - Attribuer un poids à chaque moteurs (en fonction de la pertinence par exemple).
- Solution idéale est une combinaison des différentes méthodes.

Méta-moteurs (6/6) - Problèmes

- Utiliser plusieurs moteurs n'élimine pas les problèmes liés au principe même des moteurs :
  - Indexer tous les documents.
  - Formuler une requête.
  - Pertinence est un résultat statistique.
- Tendance accrue à proposer trop ou trop peu de documents.
- Coûts peuvent devenir très important si le nombre de sources augmente.

Contenu

- Introduction
- Représentation
- Traitement
- Recherche
- Améliorations
- Méta-moteurs
- Référencement
- Conclusions

Référencement (1/11) - Contenu

- Actif
- Passif
- Mots-clés
- Positionnement
- Keyword Marketing

Référencement (2/11) - Actif

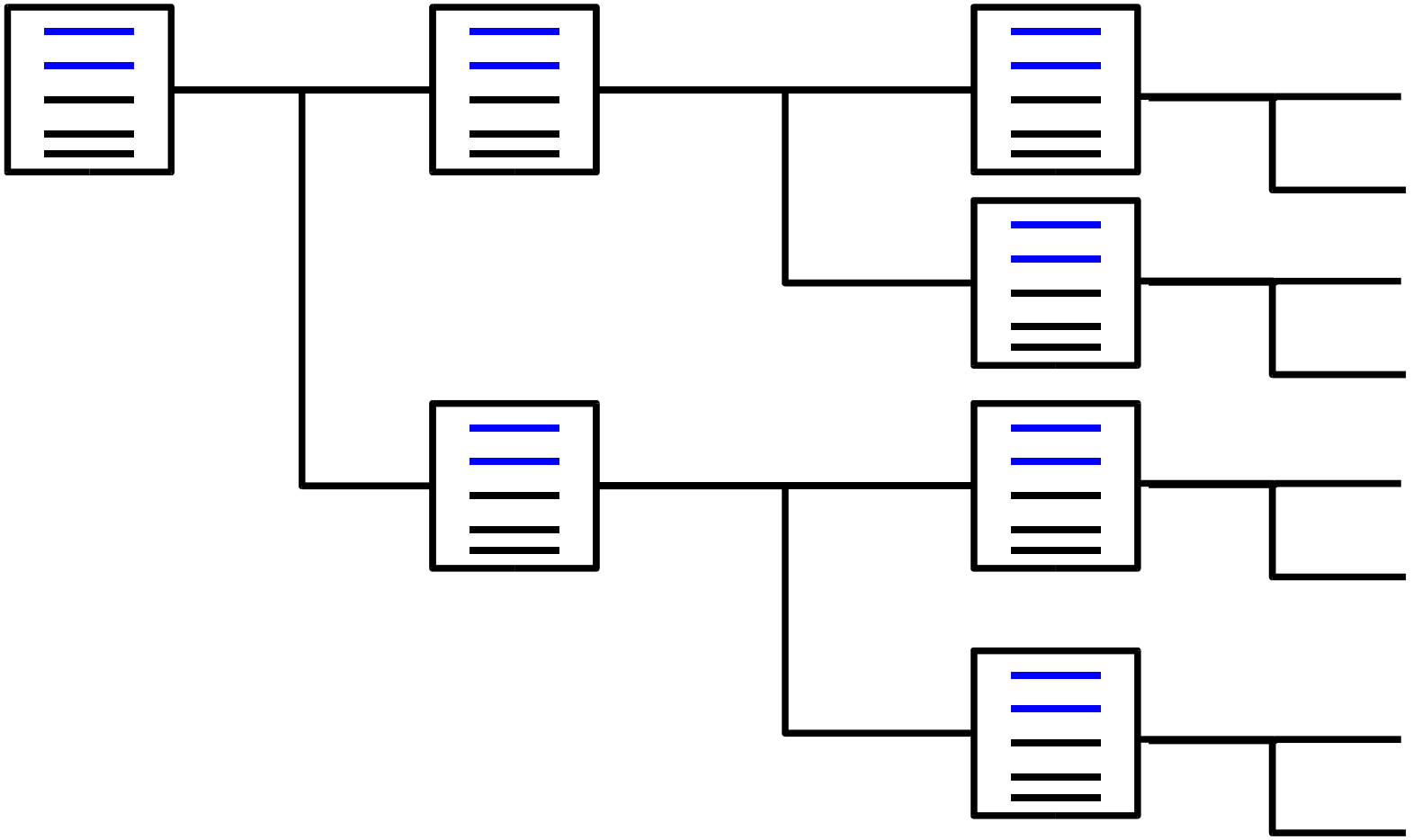
- Démarche active des concepteurs de sites.
- Contacter les moteurs de recherche.
- Se référencer manuellement via l'Open Directory Project.

Référencement (3/11) - Passif

- Attendre d'être trouvé par un robot.
- Robots sont logiciels qui scannent Internet à la recherche de nouveaux documents.
- Partant d'un ensemble de documents, les robots suivent les hyperliens de manière récursive :
  1. Soit en utilisant les premiers liens de tous les documents (méthode "depth-first").
  2. Soit en utilisant tous les liens mais en s'arrêtant à un niveau de profondeur donné (méthode "breadth-first").

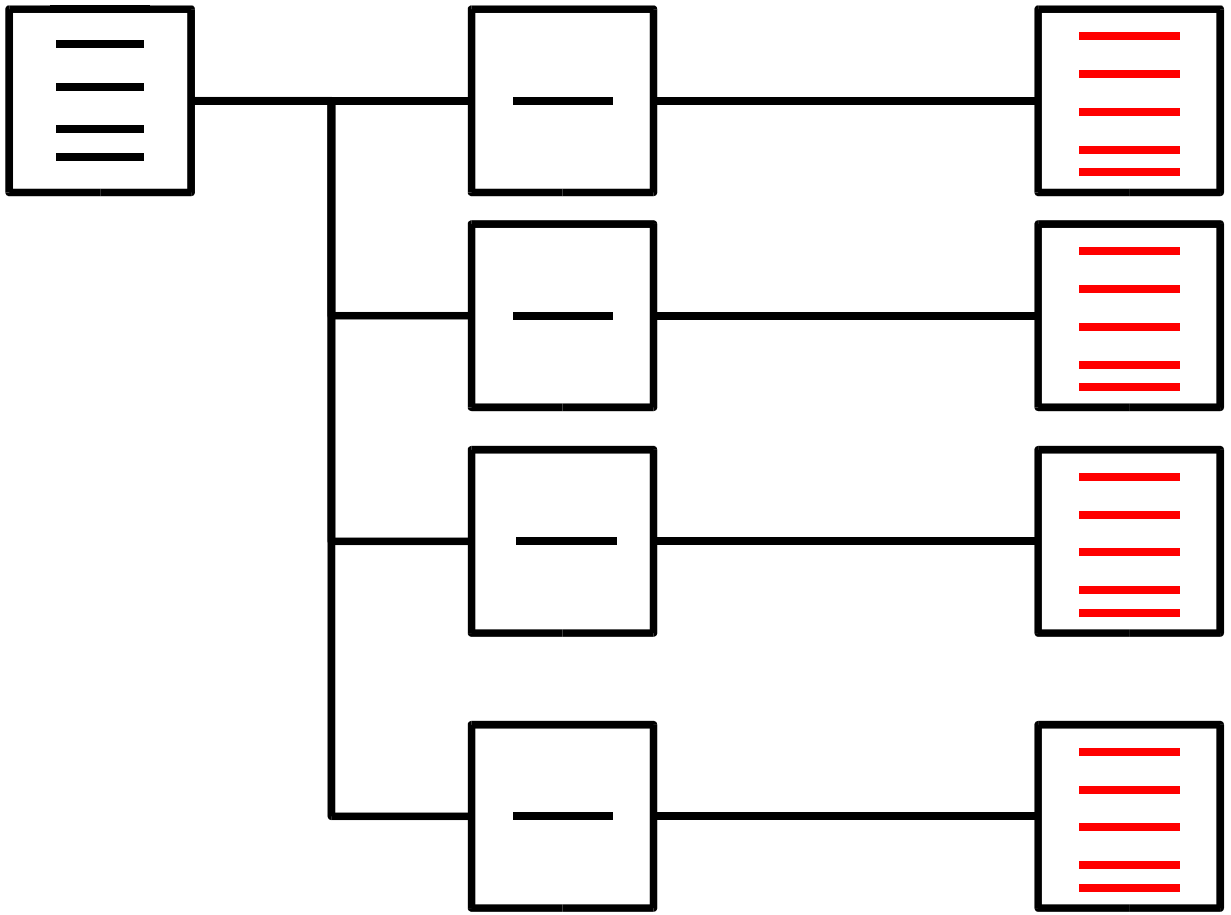
Référencement (4/11) - Depth-First

Prendre Premiers Liens (ex: 2)



Référencement (5/11) - Breadth-First

Profondeur Limitée (max: 2)



Référencement (6/11) - Mots-clés (1/2)

Situation

- Problème :
  - Requêtes sont basées sur de mots-clés ou phrases.
  - Moteur «trouve» les mots-clés dans les documents.
- Importance des mots-clés extraits :
  - Texte dans le document.
  - Méta-données (balise <meta> en html).
- Enjeux :
  - Mots-clés générique ⇒ document perdu dans masse.
  - Mots-clés précis ⇒ document jamais trouvé.

Quelques conseils

- Essayer de se mettre dans la peau d'un internaute.
- Demander un feedback de personnes.
- Trouver la liste des mots-clés et phrases les plus utilisés sur Internet.
- Spécifier la localisation (magasins, services...).
- Utiliser modificateurs pour rendre mots-clés et phrases génériques plus spécifiques.

Référencement (8/11) - Positionnement

- Chaque moteur utilise ses propres algorithmes de recherche :
  - Pertinence des mots-clés varie d'un moteur à l'autre.
  - Utilise des méthodes supplémentaire (Google et PageRank).
  - Possibilité «d'acheter» un bon positionnement.
- Bien positionner un site est donc un art et une science :
  - Connaître (ou essayer) le fonctionnement des moteurs.
  - Combiner avec des bons choix (ex: Noms de domaine).

Référencement (9/11) - Keyword Marketing (1/3)

Définition

- Keyword marketing a pour objectif de «faire trouver» les bons documents.
- Plusieurs significations :
  - Trouver les bons mots-clés (ou phrases) pour optimiser les moteurs (SEO - Search Engine Optimization).
  - Actions permettant de donner une visibilité intelligente d'un document (ex: Banners...).
- Compétitions de Googleating.

Référencement (10/11) - Keyword Marketing (2/3)

Problèmes

- Utiliser le keyword marketing pour faire monter des pages artificiellement.
- Techniques des keyword spamming :
  - Mettre des mots-clés en bas de site.
  - Ajouter du texte invisible dans un document.
- Réaction des moteurs : Développer des techniques pour éliminer le bruit.
- Sociétés de SEO essayent maintenant de trouver des accords avec les moteurs.

Référencement (11/11) - Keyword Marketing (3/3)

Éthique

- Exemples «éthiques» :
  - Ajouter une page «site map» (conseillé).
  - Créer des sites de références.
  - Ajouter des liens entre sites.
- Exemples non «éthiques» :
  - Changer le contenu d'une page en fonction de l'appelant (cloaking).
  - Créer automatiquement des pages de liens.

Contenu

- Introduction
- Représentation
- Traitement
- Recherche
- Améliorations
- Méta-moteurs
- Référencement
- Conclusions

Conclusions

- Les logiciels nécessitent une représentation des documents.
- La plupart des moteurs de recherche font un traitement particulier des documents.
- Nombreux modèles existent pour trouver les documents intéressants répondant à une requête (ex: Booléen, vectoriel, ...).
- Modèles utilisés aujourd'hui sont complexes.
- Utiliser la structure hyperliens permet d'identifier des pages ayant un rôle d'autorité.