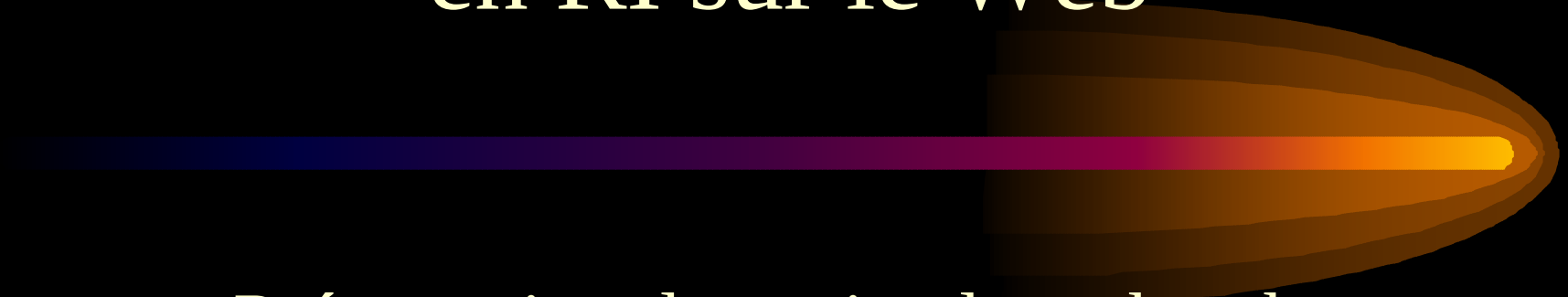
Reformulation interactive de requêtes en RI sur le Web



Présentation du projet de recherche
DIC-9410, UQAM

Emmanuel Chieze

15/05/2003

Plan de la présentation

- ***Problématique du RI sur le Web***
- Approches existantes de reformulation de requêtes
- Notre approche de la reformulation de requêtes
- Évaluation de la démarche
- Conclusion

Problématique du RI sur le Web

- Caractéristiques du RI sur le Web
  - RI ad-hoc
  - RI plein-texte : modèle du sac de mots
  - Rapprochement requêtes-documents par filtrage booléen et tri vectoriel ou autre
  - privilégie la précision sur les premiers documents (utilisation standard)

Problématique du RI sur le Web

- Caractéristiques fondamentales du Web comme collection de documents
  - collection gigantesque et multilingue
  - hétérogénéïté des domaines, genres et styles
  - => espace de recherche qualitativement différent des collections contrôlées (Blair, 2002)

Problématique du RI sur le Web

- Caractéristiques des requêtes sur le Web
  - devraient être
    - très descriptives
    - très discriminatoires
  - mais sont
    - extrêmement courtes (2 mots en moyenne)
    - et n'utilisent presque pas les fonctionnalités de
      - spécification d'expression
      - spécification de termes à exclure

Problématique du RI sur le Web

- Conséquences
  - la formulation initiale de la requête peu rarement conduire à de bons résultats
  - un processus de reformulation est indispensable pour aboutir à une précision satisfaisante
  - l'utilisateur seul ne le fait généralement pas
  - il doit donc être guidé dans ce processus

Plan de la présentation

- Problématique du RI sur le Web
- ***Approches existantes de reformulation de requêtes***
- Notre approche de la reformulation de requêtes
- Évaluation de la démarche
- Conclusion

Approches de reformulation

- Approches automatiques :
  - supposent une requête assez précise pour fonctionner, où les termes se désambiguisent mutuellement
  - condition rarement obtenue sur le Web
  - exemple : cours de cuisine
    - tu cours à la cuisine
    - cours de cuisine à l'ITHQ
- Nécessité d'une approche interactive

Approches de reformulation

- Approches interactives : rétropropagation de la pertinence
  - nouvelle requête = combinaison linéaire de la requête initiale et des premiers documents pertinents de son résultat
  - ne s'applique pas au modèle booléen
  - problèmes fondamentaux :
    - Quelles dimensions de la pertinence sont privilégiées dans ce processus ?
    - Changement de paradigme et approche de la boîte noire

Approches de reformulation

- Approches interactives : enrichissement de la collection par utilisation de thésaurus externe
  - problème de l'adéquation du thésaurus à la collection de documents (choix du vocabulaire, des liens entre termes)
  - problème de sa mise à jour (évolution de la langue, noms propres)
  - problème de l'interaction avec l'utilisateur : risque de présenter trop de choix d'hyperonymes, de quasi-synonymes ...

Approches de reformulation

- Approches interactives : sélection de termes additionnels à partir des documents du résultat initial (1)
  - peu de recherches dans ce sens
  - Belkin et al. (2000) :
    - évaluation des documents par l'utilisateur (changement de paradigme)
    - extraction probabiliste de termes des docs pertinents
    - évaluation des termes par l'utilisateur
    - enrichissement de la requête

Approches de reformulation

- Approches interactives : sélection de termes additionnels à partir des documents du résultat initial (2)
  - variante de l'analyse locale du contexte (Xu et Croft, 1996) :
    - extraction statistique de termes en co-occurrence maximale avec ceux de la requête initiale
    - évaluation des termes par l'utilisateur
    - enrichissement de la requête
    - mais repose sur des paramètres globaux
    - ne tient pas compte de la collocation

Plan de la présentation

- Problématique du RI sur le Web
- Approches existantes de reformulation de requêtes
- ***Notre approche de la reformulation de requêtes***
- Évaluation de la démarche
- Conclusion

Notre approche

- Approche interactive
- Basée sur la sélection de termes additionnels à partir des seuls documents du résultat initial
  - Mais tient compte de la syntaxe du syntagme nominal (et non de la seule collocation)
  - => Approche à paramétriser pour chaque langue traitée
- Suivie d'une évaluation des expressions par l'utilisateur
- Et d'un enrichissement de la requête
- Approche itérative

Notre approche

- Approche respectant des contraintes additionnelles liées au Web :
  - approche fonctionnant par-dessus l'utilisation d'un moteur de recherche existant, et indépendante dans les principes d'un moteur spécifique
  - approche ne nécessitant pas le rapatriement des documents initiaux si le moteur affiche des extraits pertinents
  - approche conçue pour les requêtes très courtes et donc sous-spécifiées du Web

Notre approche

- Approche d'élagage des résultats initiaux de la requête pour pallier l'absence de désambiguïsation réciproque des termes
- Approche centrée sur les termes de la requête
  - on étudie des passages des documents initiaux employant des termes de la requête
  - pour en trouver des emplois statistiquement significatifs (dans les premiers documents)
  - et trouver des termes additionnels en collocation étroite avec ceux de la requête

Notre approche

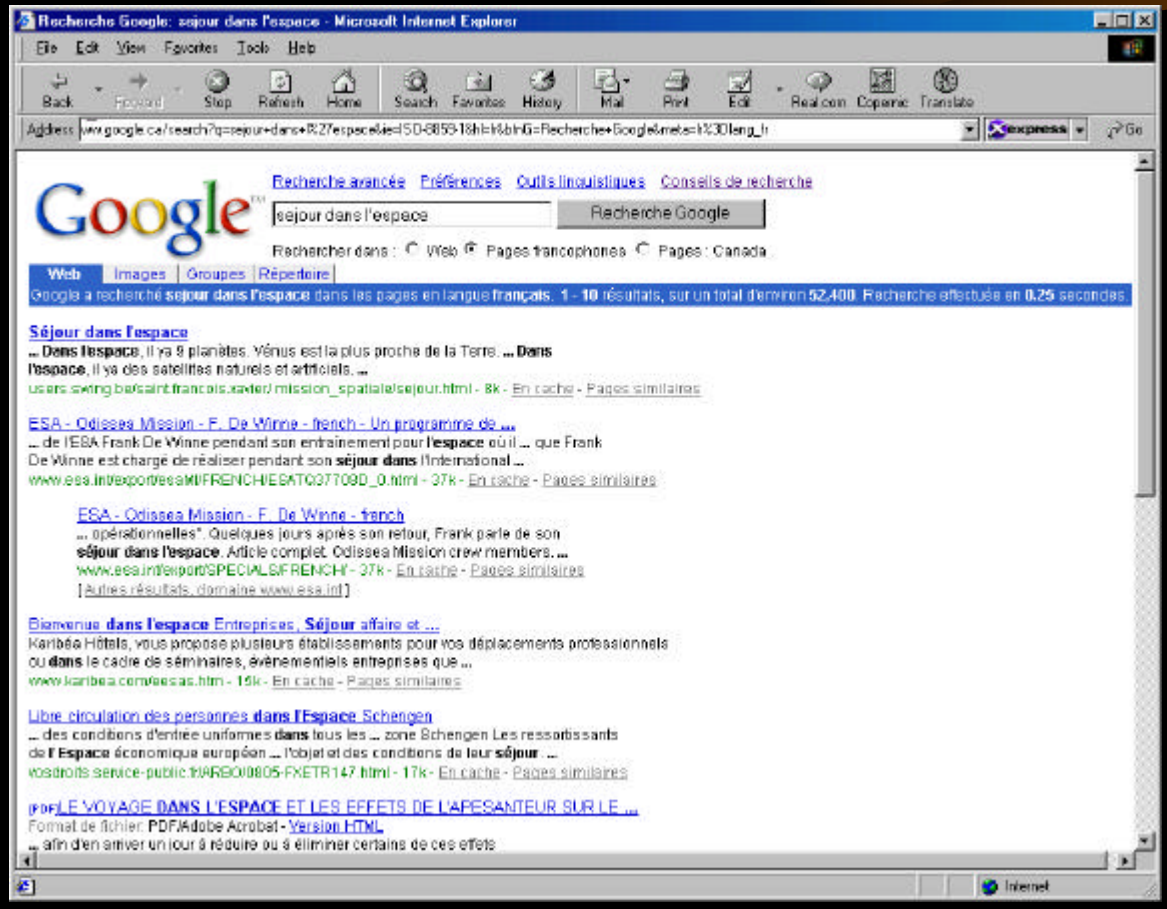
- Motivation linguistique
  - détection de cas d'homonymie : (un) livre / (une) livre
  - détection de cas de polysémie : journal (quotidien) / journal (intime)
  - sélection d'une facette particulière d'un mot : livre (objet) / livre (contenu)
  - ajout éventuel de termes associés sémantiquement (hypéronymes, hyponymes, holonymes, quasi-synonymes …)
  - prise en compte des noms propres

Notre approche

- Motivation informationnelle
  - approche en contexte permettant de limiter la détection de ces phénomènes aux cas réellement utiles pour le RI, i.e. pour compléter explicitement la désambiguïsation implicite des termes de la requête
  - en effet, une modification de la requête doit avoir un impact visible par l'utilisateur

Notre approche

- Étape 1 : exécution de la requête initiale



Notre approche

- Étape 2a : découpage des titres et extraits des N premiers documents du résultat (exple : N = 100) selon les marques de ponctuation

No Doc	Passage
1	Séjour dans l'espace
1	Dans l'espace
1	il ya 9 planètes
1	Vénus est la plus proche de la Terre
1	Dans l'espace
1	il ya des satellites naturels et artificiels
2	ESA
2	Odissea Mission
2	...

Notre approche

- Étape 2b : rattachement des articles, prépositions et conjonctions aux termes les suivant (récit de-la-journée, récit du-jour, récit quotidien)
- Étape 2c : décompte récursif de toutes les sous-chaînes de mots représentées
  - Exemple :
    - il y a des-satellites naturels et-artificiels 1
    - il y a des-satellites naturels 1
    - y a des-satellites naturels et-artificiels 1
    - il y a des-satellites 1 ...

Notre approche

- Étape 2d : élimination des sous-chaînes non-autonomes (I.e. de même cardinalité qu'une expression les contenant)
  - Exemple :
    - il y a des-satellites naturels et-artificiels 1
    - satellites naturels et artificiels 5
- Étape 2e : élimination des sous-chaines à faible fréquence (hapax et autres)
  - une fréquence > 5 indique qu'il s'agit probablement d'une expression, et que son inclusion aura un impact significatif sur la requête

Notre approche

- Étape 3 : évaluation des expressions par l'utilisateur. 3 choix possibles :
  - expression en lien direct avec le besoin en information
    - séjour dans l'espace, apesanteur ...
  - expression incompatible avec le besoin en information
    - espace schengen
  - expression neutre
    - européen, France
  - expressions présentées selon un ordre favorisant l'emploi des termes de la requête, la taille et la fréquence

Notre approche

- Étape 4 : inclusion des expressions dans la requête
  - expressions neutres ignorées
  - expressions négatives incluses par des -
    - sejour dans l'espace -"espace schengen"
  - expressions positives : plusieurs cas possibles
    - expressions de un ou plusieurs termes originaux remplacent les termes originaux : séjour-dans-l-espace
    - expressions ne comportant aucun terme original ajoutés à une section AUTRE : séjour-dans-l-espace mission OR navette

Notre approche

- Étape 4 : inclusion des expressions dans la requête
  - Difficulté de cette approche : les expressions peuvent ne pas être associées à la bonne combinaison de termes
  - Exemple : requête *projet trois gorges*
  - expressions isolées : projet-des-trois-gorges, barrage-des-trois-gorges, trois-gorges-en-chine
  - algorithme donne : *projet-des-trois-gorges barrage-des-trois-gorges OR trois-gorges-en-chine*
  - mais *projet-des-trois-gorges OR barrage-des-trois-gorges OR trois-gorges-en-chine* serait préférable du point de vue informationnel

Notre approche

- Itération : précautions additionnelles prises pour faciliter l'interaction avec l'utilisateur
  - on ne présente une expression donnée qu'une seule fois à l'utilisateur au cours du processus
  - l'évaluation positive ou négative d'une expression peut conduire à l'élimination de sur-expressions ou sous-expressions de la liste des expressions présentées à l'utilisateur

Notre approche

- Requêtes utilisées pour tester l'algorithme
  - séjour dans l'espace
  - prince-albert
  - histoire du kilt
  - projet des trois gorges
  - cours de dessin à montréal
  - petit livre rouge
  - exportation de sucre de cuba
  - hibiscus
  - sécurité dans les aéroports
  - impact des uv sur l'oeil

Plan de la présentation

- Problématique du RI sur le Web
- Approches existantes de reformulation de requêtes
- Notre approche de la reformulation de requêtes
- ***Évaluation de la démarche***
- Conclusion

Évaluation de la démarche

- Cadre expérimental
  - interface mode DOS développée en Perl
  - utilisation de Google
  - Avantages de Google
    - moteur reconnu comme le plus efficace actuellement (couverture, précision)
    - extraits significatifs => évite le rapatriement de documents
    - 100 docs/page de résultats => une seule page à télécharger
    - accès à la version indexée du doc original => facilite l'évaluation
  - Inconvénients de Google
    - limite de 10 termes par requête
    - absence de mécanisme cohérent d'exclusion d'expressions

Évaluation de la démarche

- Démarche d'évaluation
  - utilisation de la précision à 10 documents
    - précision à X docs seule mesure disponible sur le Web
    - X = 10 modélise bien l'utilisateur standard
  - nécessite une centaine de requêtes (Buckley et Voorhees, 2000) pour limiter les "ties"
    - requêtes courtes, de type RI et EI, syntaxiquement valides
    - si possible issues de TREC ou de CLEF et traduites
    - ou conçues selon ce modèle : besoin en information précis associé à chaque requête

Évaluation de la démarche

- Démarche d'évaluation
  - repose sur une notion de pertinence
    - binaire : simplification extrême mais courante de la réalité pour des requêtes de type RI, OK pour EI
    - "locale" : document pertinent s'il comporte un paragraphe contenant une occurrence des termes de la requête et ayant un sens compatible avec le besoin en information
    - thématique : on ne tient pas compte des dimensions
      - cognitives : nouveauté de l'info d'un doc par rapport aux précédents
      - situationnelles et socio-cognitives : domaine, genre et style du document, qualité de la source, adéquation du doc au problème à résoudre … (Cosijn et Ingwersen, 2000).

Évaluation de la démarche

- Démarche d'évaluation
  - évaluation statistique des résultats par comparaison de la précision à 10 docs de l'exécution initiale (référence) et de l'exécution finale (démarche proposée)
  - et utilisation du test de Wilcoxon

Évaluation de la démarche

- Évaluations complémentaires
  - validation de la formation syntaxique et de l'autonomie sémantique des expressions détectées
  - mesure d'efficacité : nombre d'itérations nécessaires pour cerner le besoin en information
  - étude des modifications apportées : spécification d'expressions, ajout de termes à inclure, à exclure
  - examen des documents ne comportant aucune expression détectée afin d'identifier d'autres facteurs permettant de présumer de leur (non-)pertinence

Conclusion

- Problématique du RI sur le Web
- Approches existantes de reformulation de requêtes
- Notre approche de la reformulation de requêtes
- Évaluation de la démarche
- ***Conclusion***

Conclusion

- Apport de ce projet de recherche
  - peu d'études publiées sur la reformulation de requêtes par ajout de termes ciblés
  - peu d'études prenant en compte la rétroaction mixte
  - peu d'études basées sur le modèle mixte filtre booléen / tri vectoriel utilisé par les moteurs de recherche
  - peu d'études basées sur le Web
  - peu d'études sur le RI en français

Conclusion

- Calendrier (à titre indicatif)
  - 05 à 06/2003 : finalisation du processus de reformulation de requêtes et conception du jeu de requêtes
  - 07 à 10/2003 : exécution des requêtes, évaluation de la pertinence des résultats et analyse statistique
  - 11/2004 : dépôt de la thèse
- Projet mené sous la co-direction de
  - Lorne Bouchard (Informatique)
  - Louisette Emirkanian (Linguistique)

Conclusion

- Références

Belkin N.J. et al. 2000. "Relevance Feedback versus Local Context Analysis as Term Suggestion Devices: Rutgers' TREC-8 Interactive Track Experience", *Proceedings of TREC-8, Gaithersburg (Maryland)*, p. 565-574.

Blair, D. C. 2002. "The challenge of commercial document retrieval, Part I: Major issues, and a framework based on search exhaustivity, determinacy of representation and document collection size", *Information Processing and Management*, 38:273-291.

Cosijn, E. et Ingwersen, P. 2000. "Dimensions of relevance", *Information Processing and Management*, 36:533-550.

Buckley, C. et Voorhees, E.M. 2000. "Evaluating Evaluation Measure Stability", *Proceedings ACM-SIGIR, New-York*, p. 33-40.

Xu, J. et Croft, W.B. 1996. "Query expansion using local and global document analysis", *Proceedings ACM-SIGIR, Zurich*, p. 4-11.