Le domaine de recherche d'information – Un survol d'une longue histoire

Jian-Yun Nie
Département d'informatique et recherche opérationnelle
Université de Montréal

Il y a 10 ans, j'ai souvent répondu aux questions sur ce que la recherche d'information était, et les différences entre les systèmes de bases de données et la recherche d'information. On aurait pensé qu'aujourd'hui, ces questions arrivent moins souvent à cause de la popularité des engins de recherche. Or, ce n'est pas le cas. Ce sont les même questions que les gens se posent, sauf qu'il est plus facile à leur expliquer en prenant les engins de recherche comme exemple. Derrière ce fait se cache la méconnaissance du domaine de recherche d'information (RI) encore aujourd'hui. Or, ce domaine a une longue histoire de 50 ans. Dans ce qui suit, je voudrais retracer très brièvement l'histoire de la RI, et les grands aspects étudiés dans ce domaine.

La naissance

Le domaine de recherche d'information remonte au début des années 1950, peu après l'invention des ordinateurs. Comme plusieurs autres domaines informatiques, les pionniers de l'époque étaient enthousiastes à utiliser l'ordinateur pour automatiser la recherche des informations, qui dépassaient la capacité humaine : il y avait une explosion d'information après la deuxième guerre mondiale.
Le nom de « recherche d'information » (information retrieval) fut donné par Calvin N. Mooers en 1948 pour la première fois quand il travaillait sur son mémoire de maîtrise [MOO 48]. La première conférence dédiée à ce thème – International Conference on Scientific Information - s'est tenue en 1958 à Washington. On y comptait les pionniers du domaine, notamment, Cyril Cleverdon, Brian Campbell Vickery, Peter Luhn, etc.

Les premiers problèmes qui intéressaient les chercheurs portaient sur l'indexation des documents afin de les retrouver. Déjà à la « International Conference on Scientific Information », Luhn avait fait une démonstration de son système d'indexation KWIC qui sélectionnait les indexs selon la fréquence des mots dans les documents, et filtrait des mots vides de sens en employant des « stoplistes ». C'est à cette période que le domaine de RI est né.

Expérimentations – une longue tradition

Dès les premiers travaux, l'aspect d'expérimentation occupait une place particulière. Pour n'importe quelle méthode, on voulait la tester expérimentalement afin de connaître son effet en réalité. Cette tradition bien ancrée dans la communauté de RI a l'avantage de se prémunir des spéculations, mais elle a aussi souvent teinté la communauté de pragmatisme.

Traçons quelques grands projets d'expérimentations dans l'histoire de la RI.

Projet Cranfield (dirigé par Cyril Cleverdon, 1957-1967) [CLE 67]

Dans la première phase de ce projet, on visait à tester l'efficacité de différentes façons d'indexer et de rechercher des documents. Ces tests sont vigoureusement contrôlés. Une collection de test est constituée d'un ensemble d'articles (18 000 dans Cranfield I) et un ensemble (1 200) de

requêtes. Ces requêtes sont évaluées par des experts afin de déterminer les réponses souhaitées - les articles pertinents. Les résultats d'une recherche automatique sont comparés avec les réponses souhaitées pour mesurer la performance en terme de précision et rappel.

Le projet Cranfield a une influence marquante sur toute l'histoire de la RI. On utilise encore aujourd'hui les mêmes principes d'évaluation pour les systèmes de RI.

Projet MEDLARS – MEDical Literature Analysis and Retrieval System (F. Wilfrid Lancaster, complete en 1968) [LAN 68]
Comme l'indique son nom, les documents dans la collection sont dans le domaine biomédical. Ces documents sont indexés manuellement, avec un vocabulaire contrôlé. Les résultats sont évalués en terme de précision et de rappel. Les résultats de ce projet montrent qu'en utilisant une approche automatique, il est possible d'atteindre la même performance avec une indexation manuelle et un vocabulaire contrôlé. Une analyse des résultats a aussi montré que l'utilisation d'un vocabulaire contrôlé et de l'indexation manuelle étaient largement responsable des cas d'échecs dans la recherche de documents pertinents, qui peuvent être évités par l'approche automatique.

SMART (Gerard Salton, 1ière version 1961-1965) [SAL 71]
Dans ce projet, une série d'expérimentations a été menée, portant sur divers sujets comme :

- la comparaison entre l'indexation manuelle et l'indexation automatique ;
- le problème de recherche d'information interactive et la rétroaction de pertinence (relevance feedback);
- l'architecture de système de RI ;
- l'utilisation du modèle vectoriel ;
- le regroupement de documents (ou clustering) ;
- etc.

Le système SMART fut réécrit dans les années 1970 et 1980 par E. Fox et C. Buckley. Ce système a été, et est encore, utilisé par de nombreux chercheurs pour des expérimentations en RI. Le système SMART est sans doute le système qui a eu le plus grand impact sur l'histoire de la RI.

Projet STAIRS - STorage And Information Retrieval System (Blair et Maron) [BLA 85]
Les documents sont dans le domaine de droit. L'indexation automatique utilise la troncation de suffixes, et la recherche exploite une liste de synonymes. Contrairement aux expérimentations antérieures qui utilisaient de petites collections, les tests de Blair et Maron porte sur une collection de taille réaliste – 40 000 documents totalisant 350 000 pages. Le résultat montre que la performance de moins de 20% de rappel est insuffisante dans le domaine de droit pour lequel le rappel est très important.

TREC - Text REtrieval Conference, (D. Harman, 1992 -) [HAR 92]
Cette série de conférences a pour objectif de tester des méthodes et des systèmes de RI avec des collections de plus grandes tailles. Elle est organisée annuellement. Les tâches (tracks) changent d'une année sur l'autre, mais elles reflètent bien les intérêts des chercheurs et les besoins réels. Au fil des années, il y a eu la RI ad hoc (la tâche classique de RI – soumettre des requêtes sur une collection statique), le filtrage de l'information, la RI non anglais (en espagnol, français, chinois) et translinguistique (retrouver des documents dans une langue différente de celle de la requête), la question-réponse, la RI multimédia (vidéo et parole), etc. Ces conférences attirent chaque année des chercheurs universitaires et industriels. Les conférences TREC ont grandement contribué au développement récent de la RI, en fournissant des collections de tests réalistes, et en offrant une nouvelle méthodologie d'évaluation. Elles ont grandement stimulé le domaine de RI.

Techniques d'indexation - statistiques et traitement de langue naturelle

L'utilisation des index remonte au 15ième siècle, peu après que l'imprimerie fut inventée. Depuis lors, ils sont utilisés notamment pour des livres et dans des bibliothèques. Les index[1] (ou termes d'indexation, termes) jouent un rôle primordial dans la RI. Ils déterminent avec quels mots on peut retrouver un document.

Le premier problème dans l'indexation fut de déterminer les éléments que l'on doit choisir comme index. Dans les premiers projets sur la RI, on s'interrogeait sur les questions suivantes :

- indexation manuelle ou automatique ?
- vocabulaire libre ou contrôlé?
- quels mots à ajouter dans la stopliste ?
- quelles méthodes de troncature ?

La première approche à l'indexation automatique KWIC ou Keyword in Context, fut introduite à International Conference on Scientific Information (ICSI) en 1958 par Luhn. On s'aperçut très tôt (dans le projet Cranfield) que les mots vides de sens (ceux de stoplistes) devrait être systématiquement éliminés. La fréquence d'occurrence de mots a été utilisée comme le critère de sélection d'index.

Il fut ensuite question de pondérer les index. Les méthodes statistiques furent exploitées dès le début de la RI. Luhn proposa une méthode basée sur la fréquence de termes dans le document (term frequency –tf). Plus tard, cette mesure a été étendue à tf*idf (idf – inversed document frequency) afin de tenir compte de la spécificité d'un terme pour un document. D'autres méthodes de pondération, telle que 2-Poisson, se sont développées plus tard (voir chapitre 1).

Dans la sélection et la pondération des index, deux aspects ressortent : spécificité (est-ce que les index sont spécifiques à un document ?) et exhaustivité (est-ce que les index couvrent tout le contenu d'un document ?). Une bonne méthode d'indexation doit faire un compromis entre ces deux aspects. Cet équilibre se retrouve notamment dans le schéma de pondération tf*idf.

Dans les premières études en RI, les techniques de traitement de la langue naturelle utilisées se sont limitées à l'analyse morphologique simple, plus spécifiquement pour la troncature ou la lemmatisation de mot. Durant le développement ultérieur, cependant, il y a eu souvent des questionnements sur le rôle de traitements automatique de langue naturelle (TALN) en RI, et plus spécifiquement dans l'indexation. Le débat à ce sujet était intense dans les années 1980, où d'une part, la popularité de l'intelligence artificielle incitait les chercheurs à utiliser des techniques de TALN dans la RI, et d'autre part, il manquait de tests avec des collections donnant des résultats convaincants. Cependant, il devient de plus en plus évident que les approches sans TALN s'approchent de leur limite. L'utilisation de TALN, que ce soit des techniques classiques basées sur des règles ou des techniques statistiques, peut permettre d'augmenter la performance (comme montre les études récentes utilisant des modèles de langue statistiques), et de raffiner la recherche (comme dans le cas de QA). Le chapitre 2 de ce livre donnera une analyse plus approfondie sur cette question.

[1] Certains argumentent que les index utilisés dans la RI ne sont pas de vrais index. Mais nous utilisons quand même ce terme ici qui est largement utilisé dans le domaine de RI, et qui tente de jouer un rôle similaire aux index manuels.

Modèles théoriques

Le rôle d'un modèle est d'abord de donner une signification au résultat de l'indexation. Un document est représenté par un ensemble de termes index et leurs poids. Si la plupart des chercheurs acceptent d'office qu'un terme index soit sensé représenter un concept important décrit dans un document, il existe différentes façons d'interpréter son poids. Un modèle théorique doit donner une interprétation précise à ce poids. Le modèle doit aussi interpréter les relations possibles entre les termes d'indexation. Ces deux fonctions nous amènent à la représentation d'un document. Une représentation similaire peut être créée pour une requête.

Finalement, un modèle doit déterminer la relation entre un document et une requête à partir de leurs représentations. Ceci se fait souvent avec un calcul de similarité.

Le développement des modèles en RI joue un rôle déterminant. C'est à travers les modèles que l'on voit la maturité du domaine. Dans l'histoire de la RI, on voit d'abord l'utilisation du modèle booléen à cause de sa simplicité, et le fait qu'il soit intuitif. Cependant, dans sa version pure sans pondération, ce modèle souffre de graves lacunes (voir chapitre 1 de ce livre). Ainsi, on proposait d'intégrer la pondération dans des modèles booléens étendus, par exemple, par l'utilisation de la théorie des ensembles flous [KRA 83]. Le modèle vectoriel est populaire grâce à sa capacité d'ordonner les documents retrouvés, sa robustesse et ses bonnes performances dans des tests. Il est sans doute le modèle le plus souvent utilisé en RI.

Les recherches sur les modèles probabilistes ont commencé depuis le milieu des années 1970. C.J. van Rijsbergen [RIJ 77], S. Robertson et K. Sparck Jones [ROB 76] sont parmi les pionniers à proposer des modèles probabilistes. L'intérêt sur les modèles probabilistes a pris son envolée à partir des années 1990, où les approches probabilistes se sont montrées performantes dans TREC (par exemple, le système OKAPI [ROB 94]]). De plus, le formalisme probabiliste offre un cadre pouvant mieux expliquer les problèmes de la RI. Ceci rejoint la tendance actuelle d'utiliser des modèles de langue [PON 98] pour la RI. Un autre facteur qui a contribué au succès de ce modèle est la disponibilité de masses de données pour l'entraînement des paramètres nécessaires.

Une autre tentative dans le développement des modèles vise à développer des modèles logiques. L'article de van Rijsbergen [RIJ 86] a provoqué un grand intérêt et des suivis sur les développements des formalismes logiques non classiques. L'objectif de ces recherches est d'une part, de développer un formalisme logique pouvant mieux capter la nature inférencielle des activités de recherche d'information, et d'autre part, offrir un traitement de l'incertitude adéquat. Le recueil [RIJ 98] contient un ensemble d'articles décrivant les études dans cette visée. Cependant, un problème important de ces approches était souvent la difficulté d'implantation. Malgré cela, les idées fondamentales proposées sont toujours utiles, et on peut espérer qu'une partie d'entre elles sera intégrée dans les développements actuels.

Améliorations techniques

De nombreuses études ont porté sur des améliorations possibles de techniques d'indexation et de recherche. Parmi les tentatives les plus marquantes, on retrouve notamment :

- Rétroaction de pertinence (relevance feedback) : Cette technique vise à étendre la portée de la recherche en intégrant les termes issus des documents pertinents, ou des documents en tête de la liste de réponses trouvées automatiquement.
- Expansion de requête : Cette technique vise à renforcer l'expression de la requête de l'utilisateur (qui est souvent très courte) par l'intégration des termes reliés (soit en exploitant un thésaurus, soit en utilisant un calcul basé sur des co-occurrences).

- Regroupement (clustering) des documents : Il vise à créer une structure entre les documents selon leurs similarités. Cette structure peut aider à la fois la recherche et la présentation des résultats.

Pertinence – la question de fond

En même temps qu'on implantait les premiers moyens pour indexer et chercher des documents, la question plus fondamentale s'est aussi posée sur la nature de la relation entre les documents recherchés et le besoin d'information de l'utilisateur – la notion de pertinence. Au cours des développements de la RI, différentes définitions ont été proposées. Dans [SAR 70], Saracevic a fait l'état d'un ensemble de définitions proposées avant 1970, notamment :
La *pertinence* est :

- la correspondance entre un document et une requête, une mesure de l'informativité du document à la requête ;
- un degré de relation (chevauchement, etc.) entre le document et la requête,
- un degré de surprise qu'apporte le document, en rapport avec le besoin de l'utilisateur,
- etc.

La définition unificatrice proposée par Saracevic est :

> La pertinence est *A* de *B* existant entre *C* et *D*, jugé par *E*,

où *A* est un intervalle de mesure pour le degré de pertinence (qui peut être binaire ou multivaluée comme de 1 à 10), *B* est l'aspect de la pertinence absolue, *C* un document, *D* le contexte dans lequel la pertinence est mesurée, et *E* le juge. Donc, pour Saracevis, il y a une mesure de pertinence (*A*), et une notion de pertinence absolue (*B*). Cette relation de pertinence absolue peut exister, ne pas exister, ou exister à certain degré, entre un document et le contexte de recherche. Il faut comprendre ici que le contexte *D* contient non seulement l'expression du besoin d'information qui est la requête, mais aussi tous les facteurs contextuels qui influencent le jugement de pertinence.

Bien que cette définition ne fournit pas plus de précision sur la nature de la pertinence, ni la façon dont elle doit être déterminée, elle identifie au moins les acteurs sur cette notion. Nous pouvons voir que ces acteurs sont assez nombreux et souvent mal identifiés (notamment les facteurs contextuels). Pour mieux cerner la nature de la pertinence, la pertinence est souvent étudiée dans le processus global de la recherche de l'information ou de communication, d'un point de vue cognitif. C'est notamment le point de vue adopté dans l'approche ASK de Belkin [BEL 82]. On s'est aussi questionné sur la relation logique que la pertinence implique entre le document et la requête [COO 71]. Pour des études plus récentes, on peut lire le numéro spécial du Journal of the Americal Society for Information Science [FRO 94]. Les investigations sur la pertinence ne peuvent pas éviter l'aspect l'utilisateur (ou le *juge* dans la formulation de Saracevic), ce qui rend cette notion subjective. Nous savons certainement beaucoup plus sur la pertinence que dans les années 1970. Mais nous sommes encore loin d'une réponse claire à la question : qu'est-ce que la pertinence ?

Ère Internet

Le domaine de recherche d'information fut créé à cause de *l'explosion de l'information* dans les années 1950. Ce terme ne décrit jamais aussi bien la réalité qu'aujourd'hui, dû à la popularisation de l'Internet et du Web. Cette nouvelle explosion a propulsé la RI au premier plan. Les techniques développées dans la RI sont de plus en plus en demande. Mais cette explosion apporte aussi de nouveaux problèmes auxquels la RI ne s'est jamais confrontée : collection gigantesque, dynamique et changeante, surabondance de l'information, données multimédia, données réparties, multilinguisme, et interaction entre utilisateur et le système.

- Sur le Web, on ne peut plus créer une collection statique. La collection (qui est le Web au complet) est une collection gigantesque qu'il est impossible (au moins pour le moment) de couvrir au complet. Bien que nous puissions employer des robots logiciels pour explorer automatiquement le Web à la découverte de nouveaux documents, nous ne pouvons pas affirmer que nous avons une bonne couverture de tout le Web.

- Un système de RI ou un engin de recherche retrouve en général toujours des documents, et beaucoup de documents. Parmi ces documents retrouvés, certains sont pertinents, mais noyés parmi beaucoup d'autres documents non pertinents. Plus notre collection contient des documents, plus ce problème devient aigu. Il est de plus en plus demandé que la recherche soit plus *précise*, même si on doit accepter que certains documents pertinents ne soient pas retrouvés. Cette problématique a motivé les recherches sur la question-réponse (QA) : plutôt que de proposer un ensemble de documents dans lesquels l'utilisateur peut trouver une réponse (ou l'information recherchée), on tente d'identifier directement la réponse [VOO 00]. Les recherches sur QA ont commencé il y a longtemps dans la communauté de TALN. À l'époque, on s'est heurté au problème qu'une réponse aux questions nécessitait souvent une compréhension profonde, dont la machine n'était pas capable. Ceci est toujours le problème clé à surmonter. Jusqu'à maintenant, les questions traitées portent en général sur des entités nommées pour lesquelles il existe des patrons bien typiques pour les identifier. Dès qu'on élargira les questions, le problème de compréhension pourra encore nous guetter.

- L'existence des documents non textuels (image, son, vidéo, etc.) nécessite de nouvelles façons pour les indexer et les rechercher. Les méthodes traditionnelles développées pour la RI sont surtout destinées aux textes, et ne sont pas directement applicables à d'autres médias.

- Aucun engin de recherche ne peut maintenant prétendre qu'il connaît tous les documents sur le Web. Chaque engin de recherche n'indexe qu'une partie de ce grand espace. Pour avoir une plus grande couverture de tous les documents, il est donc nécessaire de faire collaborer plusieurs engins de recherche. De cette façon, nous arrivons à un système de RI répartie (ou méta-engin de recherche).

- L'utilisation des langues différentes pose un autre problème. Avec une requête en français, on ne peut retrouver que des documents en français (à moins de coïncidences). Or, la pertinence d'un document est souvent indépendante de la langue utilisée. Ainsi, nous avons besoin d'outils pour la recherche d'information translinguistique ou multilingue. Les grands problèmes à traiter sont notamment la traduction de la requête et le regroupement des réponses en différentes langues dans une seule liste (si on utilise toujours une liste ordonnée pour présenter les résultats).

- La présentation des résultats d'une recherche est toujours (sauf quelques exceptions) faite avec une liste de documents dans l'ordre inversé de leurs similarités avec la requête. Bien que des chercheurs aient posé des questions à cette présentation, et aient proposé quelques alternatives (par exemple, présenter en groupes ou clusters), il a fallu attendre l'avènement du Web pour que l'on prenne pleinement conscience que cette présentation ne suffit plus. Beaucoup d'utilisateurs préfèrent naviguer sur le Web plutôt que simplement poser des requêtes (voir chapitre 4). La recherche des documents multimédias nécessite aussi d'autres façons d'indexer, de présenter et d'évaluer les résultats (voir par exemple, chapitres 9 et 10). L'utilisateur interagit aussi plus avec le système. Tous ces facteurs nous amènent à développer d'autres façons de présenter les résultats et de gérer les interactions avec l'utilisateur.

La francophonie de RI

Le domaine de RI est issu des milieux anglophones, surtout en Grande Bretagne et aux États-Unis. Les développements en recherche d'information dans la communauté francophone ont commencé dans les années 1980. La première conférence organisée en France dédiée à la recherche d'information est RIAO – Recherche d'Informations Assistée par Ordinateur – tenue à Grenoble en 1985. Dans les années 1980, relativement peu d'équipes de recherche francophones étaient actives en RI. L'équipe du laboratoire Génie Informatique (LGI), IMAG en était une des rares. Durant les années 1980, ce groupe a notamment développé le prototype IOTA qui utilise un schéma d'indexation statistique combinant une analyse syntaxique simple du français [CHI 87]. Il utilise aussi des connaissances expertes pour déterminer la collection la plus appropriée à interroger.

Dans les années 1990, le domaine est devenu plus actif. Plusieurs groupes de recherche ont identifié la RI comme un des thèmes importants. On voit des équipes de recherche apparaître en France, en Suisse, en Belgique, au Canada et ailleurs.

Quelques années après le lancement de TREC, l'INIST (Institut de l'information scientifique et technique, Nancy) a lancé un projet similaire – Amaryllis I (1996-1997). Ce projet vise à tester les systèmes et les approches pour des collections de documents en français. En 2002, Amaryllis s'est associé aux expérimentations de CLEF – Cross-Language Evaluation Forum, organisées par la communauté européenne.

La communauté de RI francophone a maintenant constitué une masse critique d'assez grande taille. Il est maintenant temps qu'elle joue un rôle plus actif sur la scène internationale. Cet ouvrage fait l'état des recherches effectuées dans cette communauté. Il est mon souhait profond que plus de travaux des équipes francophones apparaîtront dans l'histoire future de la RI.

Références

[BEL 82] Belkin, N. J., Oddy, R. N., Brooks, H. M., ASK for information retrieval. *Journal of Documentation*, 33(2): 61--71, June 1982.

[BLA 85] Blair, D.C., Maron, M.E., An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Commun. of the ACM*, 28, 1985, pp. 289-299.

[CHI 87] Chiaramella, Y., Defude, B., A prototype of an intelligent system for information retrieval: IOTA. *Information Processing and Management*, 23(4):285- 303, 1987.

[CLE 67] Cleverdon, C.W., The Cranfield tests on index language devices. *Aslib Proceedings* 19(6), 173-193, 1967.

[COO 71] Cooper, W. S., A definition of relevance for information retrieval. *Information Storage and Retrieval*, 7: 19-37, 1971.

[FRO 94] Froehlich, T. J. (ed.), Special topic issue on relevance research. *Journal of the American Society for Information Science*, 45(3), 1994.

[HAR 92] Harman, D. K. (ed.), NIST Special Publication 500-207: The First Text REtrieval Conference (TREC-1), 1992.

[KRA 83] Kraft, D. H., Buell, D. A., Fuzzy Sets and Generalized Boolean Retrieval Systems. *International Journal of Man-Machine Studies*, 19(1): 45-56, 1983

[LAN 68] Lancaster, F.W., *Evaluation of the MEDLARS Demand Search Service*, National Library of Medicine, Bethesda, Maryland, 1968.

[MOO 48] Mooers, C.N., *Application of Random Codes to the Gathering of Statistical Information*, MIT Master's Thesis, 1948.

[PON 98] Ponte, J.M., Croft, W.B.. A language modeling approach to information retrieval. In *Proceedings of the 21st ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98)*, 1998, pp. 275 – 281.
[RIJ 77] van Rijsbergen, C. J., A theoretical basis for the use of co-occurrence data in information retrieval. *Journal of Documentation*, 33: 106-119, 1977.
[RIJ 86] van Rijsbergen, C.J.. A non-classical logic for Information Retrieval. *The Computer Journal*, 29(6):481-485, 1986.
[RIJ 98] van Rijsbergen, C.J. Lalmas, M. Crestani, F., *Information Retrieval: Uncertainty and Logics - Advanced models for the representation and retrieval of information*, Kluwer Academic Publisher, 1998.
[ROB 76] Robertson, S., Sparck Jones, K., Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3):129--146, 1976.
[ROB 94] Robertson, S., Walker, S., Jones, S., Hancock-Beaulieu, M. M., Gatford, M., Okapi at TREC-3, *NIST Special Publication 500-225: the Third Text REtrieval Conference (TREC-3)*, pp. 109-126.
[SAL 71] Salton, G., The SMART Retrieval System. Prentice Hall, Englewood Cliffs, NJ, 1971.
[SAL 83] Salton, G., Fox, E.A., Wu, H. Extended Boolean Information Retrieval. *Communications of the ACM* 26(11), 1022-1036, 1983.
[SAR 70] Saracevic, T. The concept of "relevance" in information science: A historical review. In Saracevic, T. (Ed.), Introduction to Information Science, 111-151. New York: R.R. Bowker, 1970.
[VOO 00] Voorhees, E.M., Tice, D.M., The TREC-8 Question Answering Track evaluation, *Proceedings of the 8th Text Retrieval Conference*, NIST, 2000.