67th IFLA Council and General Conference
August 16-25, 2001

Code Number: **099-183(WS)-F**
Division Number: **VI**
Professional Group: Information Technology **Workshop**
Joint Meeting with: -
Meeting Number: 183
Simultaneous Interpretation: -

Accès multilingue aux systèmes d'information

Carol Peters
Istituto di Elaborazione della Informazione, CNR
Pisa, Italy
E-mail: carol@iei.pi.cnr.it

Páraic Sheridan
MNIS-TextWise Labs, Syracuse
New York, NY, USA
E-mail: paraic@textwise.com

Résumé :

Avec le développement rapide d'une société globale de l'information, le concept de bibliothèque a évolué, pour recouvrir toutes sortes de collections d'information, sur toutes sortes de support, et utilisant des méthodes d'accès différentes. Les usagers des réseaux actuels d'information et des bibliothèques électroniques ne sont plus limités par les frontières géographiques ou spatiales, ils veulent pouvoir trouver, retrouver et comprendre une information pertinente, où qu'elle soit et quelle qu'en soit la langue. Pour cette raison, on a donné beaucoup d'attention ces dernières années à l'étude et au développement d'outils et de technologies pour l'accès multilingue à l'information. Le cours fournira aux participants un aperçu des principales avancées dans ce secteur. Les sujets traités couvriront : le codage des caractères, les moyens spécifiques à mettre en œuvre pour des langages et des écritures particuliers, la localisation et la présentation des résultats, les techniques de recherche par croisement linguistique, l'importance des outils.

I -Introduction

La société globale de l'information a radicalement transformé la façon d'acquérir la connaissance, de la disséminer et de l'échanger, provoquant une révolution dans le monde des bibliothèques. Les utilisateurs de collections mises en réseau et distribuées au plan international ont besoin de pouvoir trouver, retrouver et comprendre une information pertinente, quelles qu'en soient la langue et la forme de stockage. Beaucoup d'utilisateurs ont une connaissance partielle des langues étrangères, mais leur capacité peut se révéler insuffisante pour formuler correctement les équations de recherche propres à leur besoin d'information. Ces utilisateurs seront considérablement aidés s'ils peuvent entrer leur requête dans leur langue maternelle, car ils sont capables d'examiner et d'extraire l'information de documents pertinents, même s'ils ne sont pas traduits. Les utilisateurs monolingues, pour leur part, pourront utiliser les aides à la traduction pour comprendre les résultats de recherche dans une autre langue que la leur.
C'est pourquoi, ces dernières années, on a porté une grande attention à l'étude et au développement de méthodologies et d'outils pour l'Accès Multilingue à l'Information et la Recherche Multilingue Simultanée. C'est un champ pluridisciplinaire complexe, dans lequel convergent les traitements du langage et les techniques de recherche. Le but du cours est de contribuer à la prise de conscience des buts poursuivis et des différents composants indispensables à la construction d'interfaces multilingues et en langages croisés pour les systèmes de bibliothèque électroniques. Ce papier donne un bref aperçu des principaux sujets couverts. Pour une approche plus détaillée, le lecteur est renvoyé à 1

2 - Traitement de texte multilingue

Dans la recherche d'information, on obtient une représentation du texte examiné en extrayant des "clefs d'index" d'une collection de documents ou du texte de la question du chercheur. Dans une approche simple, ce processus d'extraction se compose de quatre étapes fondamentales : conversion des caractères, extraction de mots (marquage), retrait des mots vides et normalisation des mots restants. Ces étapes ont été largement étudiées dans le contexte d'une recherche sur des textes anglais, mais il faut faire face à de nouveaux défis lorsque l'information est en plusieurs langues.

Reconnaissance des langues.
Puisque les traitements de textes pour extraire les clefs d'index impliquent souvent des étapes fondées sur une connaissance des spécificités de la langue, il est important d'établir en premier lieu la langue du texte, si celle-ci n'est pas connue. A ce jour, beaucoup d'approches ont été utilisées pour résoudre le problème de la détermination de la langue de textes généraux. On a essayé des approches s'appuyant sur la présence de caractères particuliers (2) dans les textes, ou de n-grammes (3) ou même la présence de certains mots (4). Une démarche fréquente pour identifier une langue dans un contexte multilingue est d'utiliser les mots vides propres à chaque langue.

Codage des caractères.
Alors que la plupart des langues européennes sont couvertes par la norme de codage ISO-8859-1 (Latin-1), les accès multilingues aux langues non latines impliquent la caractérisation des codages. Le codage d'une langue, ou plus spécifiquement le codage des jeux de caractères utilisés pour représenter l'alphabet d'une langue donnée, spécifie la relation graphique entre l'écrit et sa représentation binaire. Un codage de caractère est donc spécifique d'un alphabet donné, et dans beaucoup de cas, il existe des relations graphiques différentes pour le même alphabet (par exemple, l'alphabet cyrillique utilisé en Russie). Selon le nombre de caractères nécessaires à la graphie d'une langue, la combinaison d'encodage peut se faire sur un simple byte (par exemple l'allemand) ou peut demander un double byte (par exemple le chinois).

Afin de fournir une combinaison d'encodage unique pour toutes les langues du monde, le consortium UNICODE (www.unicode.org) a établi la norme UNICODE. Cette norme fournit un système de codage de caractères élaboré pour permettre l'échange, le traitement et l'affichage de textes écrits dans les diverses

langues du monde moderne. Dans les traitements de texte pour les accès multilingues, les systèmes conformes à UNICODE utilisent souvent les normes des bibliothèques pour convertir le codage initial des textes (par exemple Shift JIS pour le japonais) dans un format UNICODE (par exemple UTF-8) comme une seule représentation normalisée

Marquage spécifique des langues.
Une fois que la langue d'un texte est reconnue et que l'encodage des caractères a été normalisé, il faut identifier les mots spécifiques utilisés. Dans beaucoup de langues, l'opération est facile parce que les mots sont séparés par des espaces, mais dans d'autres langues, les mots sont agrégés ou concaténés pour former de nouveaux mots. Dans les cas les plus difficiles, il n'y a pas d'espace entre les mots (par exemple le japonais ou le chinois) et le marquage doit déterminer toutes les limites des mots. Dans ce cas, un dictionnaire ou un lexique des mots valides de la langue est utilisé pour trouver les mots autorisés. On examine automatiquement une phrase du texte pour trouver l'ensemble de mots du dictionnaire qui recouvre totalement l'ensemble des caractères trouvés dans le texte. Dans l'étape du marquage, la ponctuation est enlevée et des traits d'union sont introduits entre les segments de mots.

Pour réduire les clefs d'index à introduire dans la représentation du texte, les mots à faible valeur significative sont souvent supprimés, les mots dits "vides", par exemple en anglais *the* et *at*. Il est possible d'inclure de 30% à 50% des mots dans une liste de mots vides, leur élimination peut donc avoir un impact significatif sur les index. Les mots vides, dans une langue donnée, sont en général déterminés sur la base de chaque élément du discours (par exemple les déterminants, les prépositions) ou par la grande fréquence dans un texte échantillon.

Normalisation des mots.
L'étape finale dans un traitement de texte pour élaborer des index de recherche implique la normalisation des mots restants après l'élimination des mots vides. La forme la plus fréquente de normalisation consiste à réduire les mots à une forme radicale en enlevant les suffixes ou les inflexions. Dans le cas le plus simple, un algorithme lexical enlève simplement les suffixes courants (par exemple "-s", "-es", "-ation" en anglais) dans un processus itératif jusqu'à la forme la plus courte. Le meilleur algorithme de ce type a été développé par Porter (6) pour l'anglais et des algorithmes similaires ont été développés pour d'autres langues (5). Une alternative est de faire une analyse morphologique du texte plus linguistique pour identifier les racines des mots. La normalisation des mots apporte une plus grande efficacité dans les traitements de texte et dans les procédures de recherche. C'est particulièrement vrai pour les langues européennes qui ont une morphologie inflectionnelle plus riche que l'anglais.

Dans l'étape de normalisation, il est aussi fréquent, surtout dans le contexte d'un accès multilingue, d'identifier les phrases à plusieurs mots comme des clefs d'index uniques pour que ces phrases soient traduites comme une unité plutôt que comme des mots séparés... Dans de nombreux cas, une traduction mot à mot ne donne pas une bonne traduction (par exemple *fast-food* en français et en allemand). Ces phrases peuvent être identifiées par comparaison avec un dictionnaire ou un lexique de phrases connues, ou en utilisant des procédures statistiques qui reconnaissent les mots souvent co-occurrents comme des phrases potentielles.

Résumé
Les étapes suivies dans un traitement de texte multilingue dépendent des langues concernées et de l'approche globale utilisée dans un système donné pour l'accès multilingue à l'information. Les clefs d'index qui résultent de la phase de traitement du texte doivent être compatibles avec le moyen employé pour coupler les demandes dans une langue aux documents en plusieurs langues. Il est donc important de comprendre les différentes approches pour la recherche d'information par langage croisé et la nature des moyens utilisés dans chaque approche.

3 - La recherche multilingue simultanée

Fondamentalement, dans la recherche multilingue simultanée, il faut développer des méthodes qui comparent avec succès les demandes aux documents, quelle qu'en soit la langue, et qui classent les documents obtenus par ordre de pertinence. En recherche monolingue, pour classer, on utilise traditionnellement des comparaisons et pondérations de mots ; avec la recherche multilingue simultanée, s'ajoute le problème de la comparaison et de la pondération dans chaque langue. Il faut donc utiliser des outils capables de traduire de la langue de la requête vers celle du document ou vive versa, et de résoudre le problème de la polysémie, déjà présent en recherche monolingue, mais beaucoup plus important quand on relie plusieurs langues. Trois approches principales ont été expérimentées : la traduction automatique, les techniques fondées sur la connaissance (c'est-à-dire les thesaurus ou dictionnaires), les techniques fondées sur les corpus. Chacune de ces méthodes a donné des résultats prometteurs, mais chacune a aussi ses désavantages.

Traduction automatique

Une traduction automatique totale n'est pas une réponse réaliste au problème de la comparaison entre documents et requêtes, quelles que soient les langues. Le but d'un système de traduction automatique est de produire une version lisible et fiable dans la langue-cible du texte-source, alors que la recherche multilingue simultanée vise à trouver assez de similarités entre une requête dans une langue-source et un document dans une langue-cible pour pouvoir affirmer que le document répond plus ou moins au besoin d'information exprimé dans la requête. La traduction de toute une collection de documents dans une autre langue (celle de la requête) n'est pas seulement très coûteuse, elle implique aussi un nombre de tâches superflues du simple point de vue de la recherche, par exemple l'encodage de l'information linguistique, sémantique et pragmatique.

Les recherches fondées sur un système de traduction automatique se sont donc concentrées sur la traduction des requêtes plutôt que celle des documents. Cependant, les requêtes sont en général un ensemble de mots avec peu ou pas de structure syntaxique. De ce fait, on ne peut analyser grammaticalement les entrées avec un système de traduction automatique et les méthodes traditionnelles d'identification des polysémies ne peuvent s'appliquer dans des textes qui ne sont pas sémantiquement cohérents. Une traduction fidèle n'est donc ni possible, ni nécessaire. On n'a pas besoin d'une production cohérente et unique ; de fait, plusieurs traductions des termes d'une requête peuvent lui donner une extension aboutissant à l'enrichissement du résultat. Il a été prouvé que des méthodes plus simples et moins coûteuses techniquement se révèlent efficaces, et pour la traduction des requêtes, les techniques fondées sur les dictionnaires peuvent dépasser les systèmes de traduction automatique (7).

Les thésaurus multilingues

Les premières expériences ont montré que des thésaurus multilingues peuvent donner des résultats acceptables pour la recherche multilingue simultanée, et des systèmes basés sur les thésaurus sont désormais en vente. Un thésaurus multilingue utilisant un vocabulaire contrôlé pour l'indexage et la recherche peut être envisagé comme un ensemble de thésaurus monolingues qui englobent tous un même système de concept. Dans un vocabulaire contrôlé, on a un ensemble défini de concepts utilisés pour l'indexage et la recherche. De cette manière, le problème de l'ambiguïté est éliminé. Les utilisateurs peuvent employer un terme dans leur langue maternelle pour établir le concept identifiant correspondant afin de retrouver les documents dans d'autres langues. Dans les systèmes les plus simples, ceci peut se réaliser manuellement en consultant un thésaurus qui inclut pour chaque concept les termes correspondants en plusieurs langues et qui possède un index par langue. Dans les systèmes plus élaborés, la relation entre terme et descripteur sera réalisée en interne (8)

Dans l'approche fondée sur un vocabulaire contrôlé, les termes appropriés du vocabulaire doivent être assignés à chaque document de la collection. Traditionnellement, ceci était fait manuellement par des

experts du domaine. C'est très coûteux. On développe actuellement des méthodes semi-automatiques pour placer ces indicateurs. Il reste que les thésaurus sont très chers à construire, coûteux à maintenir et difficiles à mettre à jour. De plus, il n'est pas facile de former les utilisateurs pour qu'ils utilisent correctement les relations du thésaurus.

Actuellement, la tendance est passée d'une recherche en vocabulaire contrôlé à une recherche libre, même si, à de nombreux points de vue, la recherche multilingue simultanée est plus difficile à réaliser. Cela suppose en effet que chaque terme de la requête soit associé à un ensemble de termes de recherche dans les langues des textes, et si possible qu'on pondère chaque terme de recherche en fonction d'un degré d'occurrence dans un texte qui permettra de déterminer la pertinence du texte par rapport aux termes de la requête. La grande difficulté de la recherche multilingue simultanée en texte libre vient du fait qu'on utilise la langue vivante alors qu'avec un vocabulaire contrôlé, la recherche peut être dictée. D'un autre côté, le potentiel de la requête est plus important qu'avec un vocabulaire contrôlé.

Utilisation des dictionnaires
Beaucoup de systèmes de recherche multilingue simultanée en texte libre utilisent des dictionnaires bilingues informatisés comme moyen de transfert. Ces outils sont de plus en plus disponibles commercialement et en ligne. Comme ils ont été préparés pour un usage humain, ils demandent des traitements préalables pour être utilisés par des machines. Ceci implique essentiellement un formatage pour identifier les différentes informations lexicales : en-tête, éléments du discours, paragraphes, équivalence de traduction, etc.

On a montré qu'une simple traduction des requêtes, fondée sur un dictionnaire, dans laquelle chaque terme ou phrase de la requête est remplacé par la liste de toutes les traductions possibles, représente un premier pas acceptable vers la recherche multilingue simultanée, même si ces méthodes relativement simples ont des résultats inférieurs à ceux d'une recherche monolingue. Une traduction automatique par dictionnaire bilingue informatisé des requêtes aboutit à une baisse des résultats de l'ordre de 40 à 60% par rapport à une recherche monolingue (9) (10). Il y a trois raisons principales à cela : (i) les dictionnaires généraux n'incluent pas normalement des vocabulaires spécialisés ; (ii) échec de la traduction de termes en plusieurs mots ; (iii) problème de l'ambiguïté.

Le problème majeur dans l'utilisation des dictionnaires bilingues informatisés est peut-être l'ambiguïté. Dans les dictionnaires de traduction mot à mot, chaque mot est remplacé par tous ses équivalents traduits possibles. Quand le terme de la requête est polysémique, donc ambigu en lui-même, cela peut donner un très large ensemble de termes-cibles de recherche, dont beaucoup seront parasites et contribueront à ramener des textes non pertinents. Lorsqu'on traduit une phrase ou un document, le contexte donne des informations qui peuvent servir à lever les ambiguïtés. La brièveté d'une requête implique un manque de contexte qui empêche cette opération. Les travaux de recherche en cours portent une très grande attention à cette question.

On a prouvé que les deux méthodes, syntaxique et statistique, pouvaient réduire sensiblement les effets de l'ambiguïté et porter l'efficacité de la recherche multilingue simultanée à un niveau proche de celui de la recherche monolingue. Des requêtes bien formulées peuvent être délimitées par des limiteurs d'éléments du discours afin d'éliminer les homonymes grammaticaux et de réduire ainsi le nombre de termes-cibles incorrects générés par le dictionnaire. En particulier, les techniques d'expansion des requêtes ont montré leur efficacité pour réduire l'ambiguïté. Fondamentalement, ces techniques ajoutent de nouveaux termes, choisis d'après des critères définis, pour rendre la requête plus précise. Voir, par exemple (11) et (12)

Les techniques fondées sur les corpus.
L'approche fondée sur les corpus analyse de grandes collections de textes sur une base statistique et extrait automatiquement l'information nécessaire à la construction de techniques de traduction propres à une

application. Les collections analysées peuvent être constituées de textes parallèles (traduction équivalente) ou comparables (lié à un domaine). Les démarches principales utilisant les corpus sont l'espace vectoriel et les techniques probabilistes.

Les premiers essais à partir de corpus parallèles ont porté sur des méthodes statistiques pour extraire les données d'équivalence des termes multilingues qui pourront être utilisées en entrée pour la composition lexicale des systèmes de traduction automatique. Le problème de l'utilisation de textes parallèles comme corpus d'essai est que ces corpus sont en général liés à un domaine et qu'ils sont coûteux à constituer : il est difficile de trouver des traductions déjà existantes de la bonne espèce de documents et les versions traduites coûtent cher à créer. Pour cette raison, on s'est davantage intéressé aux potentialités des corpus comparables.

Une collection de documents comparables est constituée de documents rassemblés sur la base de la similarité des sujets traités plus que sur leur équivalence en traduction. Il faut que ces documents soient similaires en genre, registre et date. L'idée qui sous-tend l'utilisation de ces corpus est que les mots utilisés pour décrire un sujet particulier seront liés sémantiquement à travers les langues

La stratégie la plus connue de recherche multilingue simultanée fondée sur des corpus comparables est celle des thésaurus de similarité multilingue. (13) donne les résultats obtenus à partir d'un corpus de référence créé en regroupant les nouvelles d'une agence de presse suisse (SDA) en allemand et en italien par sujet et par date, puis en les fusionnant pour créer le "thésaurus de similarité. Des requêtes en allemand ont été testées sur une importante collection de documents italiens. Les résultats de cette démarche sont prometteurs, surtout lorsqu'on utilise un ensemble spécialisé dans un domaine. (14)

Un gros désavantage des techniques fondées sur les corpus est qu'elles tendent à être très dépendantes des applications. Il faut de nouveaux corpus de référence pour tout nouveau domaine.

Résumé
D'après l'état de l'art actuel, toutes les approches citées, si elles sont implémentées dans des systèmes bien organisés, testés et paramétrés, peuvent aboutir à une efficacité de l'ordre de 80% des recherches monolingues en général. Cependant, comme on a pu le voir dans ce bref survol, chaque méthode de recherche multilingue simultanée a ses limitations. Quelle que soit la méthode choisie, les outils utilisés pour donner les moyens de relier entre eux requête et collection sont un élément déterminant du succès de la recherche. Les outils préexistants, comme les dictionnaires électroniques bilingues, sont en général inadaptés ; la construction d'outils spécifiques, comme les thésaurus et les corpus de test, est coûteuse et ces outils ne sont pas totalement réutilisables. Une nouvelle application multilingue demandera la construction de nouveaux outils ou un énorme travail d'adaptation des outils existants.

On notera aussi que la plupart des systèmes en usage sont concentrés sur deux langues plutôt que sur des langues multiples. Ce n'est pas étonnant. La situation est beaucoup plus complexe quand on essaie de réaliser une recherche efficace sur plusieurs langues plutôt que sur un seul binôme ; il est nécessaire de développer une sorte de mécanisme interlinguistique, à un niveau plus ou moins conceptuel, pour permettre les transferts croisés multilingues. Dans une interlinguistique conceptuelle, les mots et les phrases de langues diverses qui réfèrent au même concept sont liés dans une représentation indépendante des langues. De cette façon, il est possible d'associer des termes équivalents dans toutes les langues et de réaliser une recherche multilingue simultanée dans toutes les combinaisons de langues, pas uniquement entre deux langues. Toutefois, la construction d'un tel outil n'est pas facile et il reste beaucoup à faire avant que nous puissions parler de vrais systèmes de recherche multilingue simultanée.

4 - Campagnes d'évaluation des systèmes de recherche multilingue simultanée

L'évaluation des systèmes joue un rôle important pour stimuler les développements. C'est particulièrement vrai pour les systèmes de recherche multilingue simultanées qui en sont encore beaucoup au stade expérimental. Il y a actuellement plusieurs réalisations internationales dans ce domaine
TREC - Text Retrieval Conference Series (http://trec.nist.gov/) qui inclut cette année une piste multilingue pour l'anglais, le français et l'arabe.

CLEF – Cross language Evaluation Forum (http://www.clef-campaign.org/). CLEF, financé par la Commission européenne, comme une partie de DELOS Network of Excellence for Digital Libraries mène une action d'évaluation pour les langues européennes et a pris la suite d'une piste pour la recherche multilingue simultanée initiée dans TREC en 1997. CLEF 20O01 a quatre missions pour l'évaluation de la recherche de texte multilingue, bilingue, spécialisée et monolingue (autre qu'anglais). La collection multilingue de cette année regroupe les corpus comparables de journaux en sis langues (hollandais, anglais, français, allemand, italien, espagnol) et des sujets en 10 langues européennes et 3 asiatiques.
NTCIR: NACSIS Test Collection for Information Retrieval (http://www.rd.nacsis.ac.jp/) est hébergé par l'Institut national pour l'informatique de Tokyo NTCIR incorpore des missions pour le chinois-anglais et le japonais-anglais

Ces activités donnent aux développeurs des forums importants pour se rencontrer, échanger des idées et des expériences, comparer des résultats. Pour les rapports sur les recherches les plus récentes en Accès Multilingue à l'Information, le lecteur est invité à se référer aux derniers comptes-rendus de ces initiatives (15), (16), (17)

Conclusion

Nous avons donné un aperçu rapide des éléments qui doivent être pris en compte quand on construit un système qui donne des fonctions de recherche et un accès à des collections de documents en plusieurs langues. Beaucoup de progrès ont été réalisés ces dernières années dans ce domaine et une dynamique importante s'est mise en place. Les efforts actuels sont centrés sur des sujets tels que la combinaison de sources multiples de correspondance de traduction pour améliorer l'association multilingue des requêtes et des documents, accès multilingue pour les langues "à faible densité" (celles pour lesquelles les outils linguistiques ne sont pas encore disponibles sur support électronique), les accès multilingues au contenu multimédia (plus particulièrement les documents parlés), et la présentation des résultats de recherches multilingues, incluant le résumé de contenu établi à travers des documents multiples en langues différentes.

Toutefois, il est évident que même si grâce à l'accès Multilingue à l'Information et à la Recherche Multilingue Simultanée la recherche a fait des progrès significatifs ces dernières années, en réalité les applications qui manipulent des documents en plusieurs langues ne donnent que des outils d'accès très simples, n'allant pas en général au-delà d'un vocabulaire de recherche contrôlé sur des champs sélectionnés, et rarement pour plus de deux langues. Une enquête récente présentée lors d'une réunion de European Digital Library projects, financée par la Commission européenne, a confirmé que si beaucoup de projets concernent actuellement les documents en plusieurs langues, très peu d'entre eux ont abouti à l'implémentation d'outils permettant la recherche sur plus d'une collection dans une langue. Il est évident qu'il reste encore à faire des efforts considérables pour transférer les résultats de la recherche dans des applications pratiques. Nous espérons que notre cours sera un pas dans cette direction.

References

1. Peters, C., Sheridan, P.: Multilingual Information Access. In M. Agosti, F. Crestani, G. Pasi (eds.). Lectures on Information Retrieval, Lecture Notes in Computer Science 1980, pp51-80, Springer Verlag, 2001.
2. Ziegler, D.: The Automatic Identification of Languages Using Linguistic Recognition Signals. PhD Thesis, State University of New York, Buffalo, 1991
3. Damashek, M.: Guaging Similarity with N-grams: Language-independent Categorization of Text. Science, Vol. 267 (No. 10) 1995
4. Souter, C., Churcher, G., Hayes, J., Johnson, S.: Natural Language Identification using Corpus-based Models. Hermes Journal of Linguistics, Vol. 13, pp. 183-203, Faculty of Modern Languages, Aarhus School of Business, Denmark, 1994
5. Wechsler, M., Sheridan, P., Schäuble, P.: Multi-Language Text Indexing for Internet Retrieval. In Proceedings of the 5th RIAO Conference, Computer-Assisted Information Searching on the Internet, Montreal, Canada, June 1997.
6. Porter, M.F.: An Algorithm for Suffix Stripping. Program, Volume 14 (No. 3), pp 130-137, 1980.
7. Ballestreros, L., Croft, W.B.: Resolving Ambiguity for Cross-language Retrieval. In Proceedings of the 20th International ACM SIGIR Conference on Research and Development in Information Retrieval, Philadelphia, PA, pp 84–91, 1997.
8. Soergel, D.: Multilingual Thesauri in Cross-Language Text and Speech Retrieval. In Working Notes of AAAI Spring Symposium on Cross-Language Text and Speech Retrieval, Stanford, CA, pp 164–170, 1997.
9. Hull, D.A., Grefenstette, G.: Querying Across Languages. A Dictionary-based Approach to Multilingual Information Retrieval. In Proc. of 19th ACM SIGIR Conference on Research and Development in Information Retrieval, Zurich, Switzerland, pp 49–57, 1996.
10. Ballesteros, L., Croft, W.B.: Dictionary-based methods for cross-lingual information retrieval. In Proceedings of the 7th International DEXA Conference on Database and Expert Systems Applications, pp 791–801, 1996
11. Ballesteros, L., Croft, W.B.: Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval. In Working Notes of AAAI Spring Symposium on Cross-Language Text and Speech Retrieval, CA, pp 1–8, 1997
12. Adriani, M., van Rijsbergen, C.J.: Term Similarity-Based Query Expansion for Cross-Language Information Retrieval. In Lecture Notes in Computer Science, 1696, 1999.
13. Sheridan, P., Ballerini, J.P.: Experiments in Multilingual Information Retrieval using the SPIDER System, In Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval, Zurich, Switzerland, pp 58-65, 1996.
14. Sheridan, P., Braschler, M., Schäuble, P.: Cross-Language Information Retrieval in a Multilingual Legal Domain. In ECDL'97 Proceedings, Pisa, Italy, pp 253–268, 1997
15. Voorhees, E.M., Harman, D.K. (eds.). The Eighth Text Retrieval Conference (TREC-8), US National Institute of Standards and Technology, 2000
16. Peters C. (ed.). Cross-Language Information Retrieval and Evaluation: Proc. of the CLEF 2000 Workshop. Lecture Notes in Computer Science, 2069, Springer Verlag, 2001
17. Kando, N., Aihara, K., Eguchi, K, Kato, H. Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization, ISBN 4-924600-89-X, 2001.