Chapitre 10

L'apport de la carte bibliographique : comparaison avec des résultats de l'ADS

L'objet de ce chapitre n'est pas de comparer les performances de la carte bibliographique à celles de l'"ADS abstract service" (ADS[1]), mais plutôt de déterminer dans quelle mesure notre système peut donner des résultats similaires et ce qu'il apporte à un système plus classique dans le cadre d'une recherche de documents.

10.1 L'expérimentation

10.1.1 Le choix de l'ADS

Le choix de l'ADS comme point de comparaison de notre système s'est fait naturellement : l'ADS est en effet le système le plus utilisé par les astronomes, mais aussi le plus complet. Il permet en effet des recherches variées (sur les auteurs, sur les termes, sur les objets astronomiques, propose des filtres par date, par revue, etc...) sur plus de 450.000 documents en astronomie et astrophysique provenant de plus de 2500 sources (journaux, actes de conférences, divers bulletins et circulaires).

La possibilité de sélectionner la revue et la période dans lesquels sont effectuées les recherches nous a permis de travailler sur le même ensemble de données : les articles de A&A publiés de mai 1994 au 10 octobre 1997 (environ 4800 documents).

10.1.2 La procédure d'interrogation

Notre expérimentation consiste à effectuer la même requête sur les deux systèmes, et d'en comparer les résultats. Mais une difficulté apparaît : les deux systèmes présentent les résultats de leurs recherches de façon très différente :

- le système de recherche de l'ADS est un système qui renvoie les documents retrouvés sous la forme d'une liste ordonnée ;
- la carte bibliographique renvoie une image permettant de localiser les classes où se trouvent des documents qui correspondent à la requête.

[1] Astrophysics Data System : service documentaire proposé par la NASA.

Ainsi, la réponse de notre système est difficile à évaluer, car elle suppose une nouvelle intervention de l'utilisateur. Celui-ci doit faire un choix (guidé par les indications que l'image lui fournit, voir 7.3) parmi les classes qui peuvent l'intéresser, avant de pouvoir accéder à une liste de documents, ordonnée elle aussi.

Afin d'éliminer l'aspect interactif, qui est une qualité de la carte bibliographique mais pose problème ici, que présente la carte bibliographique pour l'accession aux documents retrouvés, nous avons opté pour une autre manière d'effectuer les requêtes : la recherche de documents similaires à un document donné.

La recherche de documents similaires avec l'ADS

Ce type de recherche est intégré au système. A partir de tout document sélectionné, le système offre différentes méthodes pour rechercher des documents similaires. Chacune d'entre elles consiste à formuler automatiquement une requête en utilisant différents composants de l'article de départ :

1. le nom des auteurs ;
2. les mots du titre, qui sont recherchés dans les titres des autres documents ;
3. les mots du résumé, qui sont recherchés dans le résumé des autres documents ;
4. les mots-clés, qui sont recherchés dans le résumé des autres documents ;
5. une combinaison de ces composants.

C'est la quatrième méthode que nous avons utilisé : la recherche des mots-clés dans les résumés des documents. En effet, la recherche sur les auteurs est hors de notre propos, tandis que les deux autres tendent à restituer des documents très nombreux et très divers, beaucoup moins ciblés qu'avec la quatrième méthode.

La recherche de documents similaires avec la carte bibliographique

Les documents y étant classés par ressemblance, la recherche des documents similaires dans la carte bibliographique est simple : ils doivent se trouver dans la même classe que le document de départ. Ceci suppose bien sûr que ce dernier soit bien classé.

Nous ordonnons ensuite les documents de cette classe, par ordre décroissant de ressemblance avec le document de départ. Cette ressemblance est directement liée au nombre de mots-clés en commun entre le document de départ et les autres.

La taille des deux listes de documents

Nous avons limité les deux listes obtenues à un nombre avoisinant 20 documents. Nous avons en effet estimé à ce nombre les documents restitués que les utilisateurs examinent en moyenne à la suite d'une requête[2]. Nous avons donc utilisé les premiers documents des deux listes (ordonnés par ressemblance décroissante) en veillant à ne pas couper les listes entre deux documents de degré de ressemblance égal . C'est pourquoi les listes obtenues ne contiennent pas toutes exactement 20 documents.

[2]Une étude (Silverstein et al., 1998) est même plus pessimiste, puisque le nombre de 10 y est avancé.

Le protocole

Des spécialistes de différents domaines de l'astronomie ont choisi, parmi les documents de la base, des articles (les articles de départ) dans leur spécialité, sans contrainte ni influence particulière. Pour chacun des documents de départ, nous avons constitué une liste de documents potentiellement similaires à partir des documents retrouvés par les deux systèmes. Cette liste a été soumise aux spécialistes afin qu'ils évaluent la similarité des documents qu'elle contient avec celui qu'ils ont choisi auparavant.

10.2 Résultats

document de départ		nombre total	pertinents
1996A&A...313..723M	ADS	19	52%
	Carte	19	57%
	Communs	6	83%
1997A&A...322....1B	ADS	23	60%
	Carte	24	62%
	Communs	13	61%
1997A&A...317..164Z	ADS	18	22%
	Carte	13	7%
	Communs	6	16%
1996A&A...311..758S	ADS	20	55%
	Carte	18	61%
	Communs	5	60%
1997A&A...323..461H	ADS	12	83%
	Carte	12	58%
	Communs	5	100%
1995A&A...300..349M	ADS	26	46%
	Carte	20	40%
	Communs	4	100%
1996A&A...312..105C	ADS	18	72%
	Carte	19	68%
	Communs	4	100%
1994A&A...284..949G	ADS	18	66%
	Carte	12	91%
	Communs	5	100%
1995A&A...304...44L	ADS	19	73%
	Carte	18	83%
	Communs	10	90%

Tab. 10.1: *Résultats des évaluations concernant 9 documents de départ.*

Le tableau 10.1 donne les résultats de neuf évaluations. Ces évaluations sont trop peu nombreuses pour permettre d'en tirer des conclusions quantitatives sur les performances des deux systèmes, mais comme nous l'avons dit en introduction à ce chapitre, cela ne constitue pas le but de cette expérience.

Il apparaît, à la lecture du tableau 10.1, que la carte bibliographique peut donner des résultats globalement comparables à ceux de l'ADS (les moyennes du nombre de documents pertinents, pondérées par le nombre total de documents retrouvés est d'environ 60% pour les deux systèmes).

Tout aussi importante sans doute, est la signification du nombre et de la pertinence des documents retrouvés par les deux systèmes. Si ces documents sont en général souvent pertinents (la proportion des documents pertinents y est proche de 80%), ils sont également peu nombreux. En effet, ils représentent environ 50% des documents pertinents retrouvés par chaque système, qui apparaissent ainsi fournir des résultats complémentaires.

10.3 Analyse des résultats

Le cas de 1997A&A...317..164Z

Le document 1997A&A...317..164Z[3] est celui pour lequel les résultats obtenus avec les deux systèmes sont les moins bons. Ce document, qui traite de la recherche de naines brunes dans l'amas stellaire ouvert des Pleïades est défini par les mots-clés suivants : "stars : low mass,brown dwarfs", "stars : pre-main sequence", "stars : late type". Dans la carte bibliographique, nous avons retrouvé cet article dans une classe où le thème principal est la formation d'étoiles et les étoiles pré-séquence principale. Le rapport avec les naines brunes est très mince, ce qui explique le faible nombre de documents pertinents retrouvés avec la carte (en fait, un seul a été retrouvé, décrit par exactement les mêmes mots-clés que l'article de départ). A l'évidence, le mot-clé qui est à la source de l'appartenance de cet article dans sa classe est "stars : pre-main sequence", qui est non pas relatif aux naines brunes, mais aux étoiles de l'amas des Pléïades dans lequel les observations sont menées.

Une étude sur la localisation des articles décrits par le mot-clé "stars : low mass,brown dwarfs" montre que ces articles sont très dispersés sur la carte. La cause de cette dispersion est d'une part, la relativement faible fréquence d'apparition de ce mot-clé (seulement 38 fois), et d'autre part la grande diversité des mots-clés avec lesquels il se trouve associé dans la description des documents. Il se trouve aussi bien associé par exemple avec le mot-clé "stars : binaries : close", "stars : interiors", ou encore "cosmology : dark matter", et les documents concernés appartiennent aux classes différentes correspondantes. Il faudrait que de tels documents soient beaucoup plus fréquents afin que leur poids soit suffisamment important lors de l'apprentissage, pour rapprocher (sur la carte) ces différents domaines autour d'une zone relative au mot-clé "stars : low mass,brown dwarfs".

La seule méthode qui permette de retrouver les documents similaires à "1997A&A...317..164Z" est d'effectuer une requête par mot-clé (en utilisant le mot-clé relatif aux naines brunes) et d'examiner les documents sélectionnés les uns après les autres.

La localisation des documents retrouvés par l'ADS

Les résultats de notre expérimentation montrent que chacun des deux systèmes retrouve environ 50% de documents pertinents que l'autre ne retrouve pas. Se pose alors la question de savoir comment retrouver les documents manquants.

[3]Brown dwarfs in the Pleiades cluster : a CCD-based R, I survey. (ZAPATERO OSORIO M.R., REBOLO R., MARTIN E.L.)

Le cas de l'ADS est relativement simple : on peut soit formuler la requête différemment, soit examiner un nombre plus important des documents retournés. Dans les deux cas, on est obligé de contrôler un grand nombre de documents (nous avons observé qu'il n'est pas rare de devoir examiner plus de 100 documents avant de retrouver un des documents pertinents retrouvés avec la carte bibliographique).

Avec la carte bibliographique, nous avons constaté que les documents pertinents qui n'appartiennent pas à la classe du document de départ ont, dans la grande majorité des cas, des mots-clés en commun avec le document de départ. Les exceptions que nous avons constaté sont des documents qui comportent des erreurs d'indexation telles qu'une mauvaise écriture (mot-clé non reconnu) ou une mauvaise attribution des mots-clés.

Ces documents, qui partagent une fraction des mots-clés avec le document de départ, peuvent être localisés par une requête sur la carte bibliographique. Les documents les plus proches du document de départ sont repérés par la taille variable des carrés (voir 7.3). Quant au contenu des articles sélectionnés, il est résumé par les mots-clés qui les décrivent, ainsi que leur fréquence d'apparition.

10.4 Conclusion

Le système de l'ADS et la carte bibliographique sont complémentaires. Chacun permet de retrouver des documents pertinents que l'autre ne retrouve pas.

Alors que pour retrouver les documents manquants avec l'ADS, il est nécessaire soit d'examiner en profondeur la liste des documents retournés, soit d'en générer une autre (en reformulant la requête), la carte bibliographique permet la sélection de classes où se trouvent des documents proches du document de départ (à la suite d'une requête utilisant les mots-clés de ce document).

Lorsqu'un document se trouve mal classé (comme l'article traitant des naines brunes dont nous avons décrit le cas plus haut), ce qui se remarque aisément en le comparant au document moyen représentatif de la classe, une requête utilisant les mots-clés de ce document permet de repérer les classes qui contiennent les documents les plus proches. C'est d'ailleurs cette opération qui est effectuée dans les cas où la recherche ne se fait pas par rapport à un document connu, mais en fonction d'un sujet. La tâche qui suit la requête consiste alors en l'examen des documents retournés.

Les opérations effectuées sur les deux systèmes dans le but d'améliorer leurs résultats sont finalement assez semblables, puisqu'elles consistent à sélectionner davantage de documents pour y rechercher l'information voulue. L'intérêt de la carte bibliographique est que ces nouveaux documents à examiner y sont classés par thème, et résumés. A l'opposé, les documents retournés par l'ADS sont uniquement classés en fonction du nombre de termes qu'ils ont en commun avec la requête, sans aucune prise en compte des thèmes qui y sont abordés.