Première partie

Préliminaires

Chapitre 2

La recherche d'information

Dans ce chapitre, nous allons aborder les systèmes de recherche d'information d'une manière générale, ainsi que les données dont il sera question dans la majeure partie de cet ouvrage : les données textuelles.

2.1 L'information textuelle

Les données textuelles renferment plusieurs types d'information, qu'il est possible de classer en trois catégories :

- l'information structurelle : elle renseigne sur la *structure* des documents, par exemple, le nombre de paragraphes, la présence d'un sommaire, etc...
- l'information factuelle : elle renseigne sur la *valeur* de certains attributs comme la date de publication, le numéro du volume, etc...
- l'information sémantique : elle est directement liée aux mots du texte de chaque document, c'est à dire à leur *sens*.

Dans la grande majorité des cas, c'est l'information sémantique qui est recherchée. En effet, on recherche en général des documents relatifs à un thème ou un sujet donné. De plus, si l'information sémantique est la plus fréquemment recherchée, elle est également la plus difficile à caractériser car des mots différents peuvent décrire un même thème et un même mot peut décrire plusieurs thèmes. C'est ce qui explique le nombre important des travaux de recherche effectués pour développer et améliorer les systèmes de recherche de l'information sémantique dans les textes.

En résumé, dans la suite de ce document, quand il sera question d'*information dans les textes* sans plus de précision, ce sera toujours de l'information *sémantique* dont il sera question.

2.2 Aspects généraux des systèmes de recherche d'information

Il convient de définir dès maintenant ce que l'on entend par *système de recherche d'information* (SRI), car l'expression peut prêter à confusion. La définition qu'en a donnée Lancaster (Lancaster, 1968) semble bien décrire tous les travaux effectués dans le domaine. En voici à peu près la teneur : "un système de recherche d'information ne renseigne pas

ses utilisateurs sur l'objet de leurs interrogations (ne fait pas évoluer leurs connaissances), mais indique simplement quels sont les documents (s'ils existent) en rapport avec ces interrogations". Un SRI est donc un système qui facilite l'accès aux documents qui contiennent l'information recherchée, une sorte d' "index amélioré" qui prend tout son sens dans la consultation de grandes collections documentaires, dans lesquelles il est impossible de localiser efficacement de l'information sans une aide automatisée.

Un SRI est donc un logiciel que l'on utilise pour retrouver des documents relatifs à un sujet, à une interrogation.

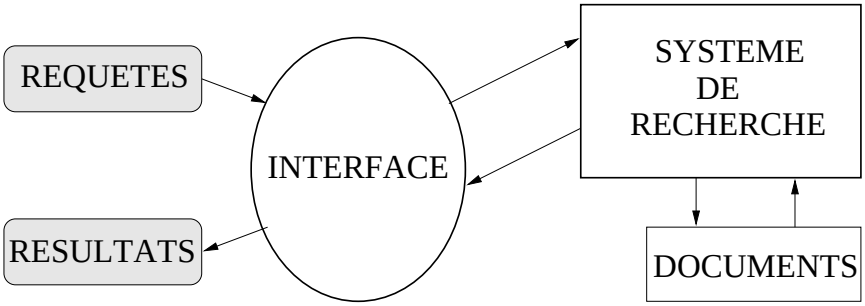

Fig. 2.1: *Système de recherche d'information typique.*

Il est donc important qu'un SRI soit en mesure de (Fig. 2.1) :

- gérer les requêtes des utilisateurs et leur communiquer les résultats des recherches ;
- accéder aux données afin d'y retrouver les informations que recherchent les utilisateurs.

Nous allons voir de plus près l'organisation des SRI autour de ces deux points, puis nous verrons comment il est possible d'évaluer leurs performances.

2.2.1 L'interface de consultation

La gestion des requêtes et la restitution des résultats sont étroitement liées à l'interface de consultation d'un SRI. C'est elle qui assure la communication entre les utilisateurs et le système de recherche proprement dit. Deux rôles incombent à l'interface : l'acquisition et l'interprétation des requêtes, et la visualisation des résultats des recherches.

2.2.1.1 Interrogation

Il existe plusieurs types d'interrogations ou langages d'interrogation :

- on peut distinguer tout d'abord l'interrogation en langage booléen, qui consiste à indiquer les termes que doivent contenir (ou non) les documents recherchés. Il est possible de combiner ces termes avec des opérateurs booléens ET, OU et NON. Par exemple, une personne intéressée par les moyens de transport au siècle dernier pourra formuler sa requête de manière à retrouver les documents qui contiennent les termes "voiture" ET "chevaux". Il se peut qu'il ne soit pas satisfait du résultat car un grand nombre des documents retournés sont relatifs au nouveau règlement sur la puissance maximale des automobiles dans les grands prix de formule-1. Il lui suffira de rechercher les documents qui contiennent "voiture" ET "chevaux" et NON "formule-1", afin d'éliminer les documents inopportuns.

L'inconvénient de ce mode d'interrogation est qu'il est basé sur un langage dont l'emploi peu être lourd : les requêtes peuvent devenir assez longues et compliquées à formuler pour une recherche élaborée.

- une autre manière d'utiliser un SRI est l'interrogation en langage libre. L'intérêt est qu'il n'est pas nécessaire d'apprendre un langage pour interroger ces systèmes. Les utilisateurs formulent leurs requêtes sans avoir à utiliser de syntaxe précise. Le système se charge de leur interprétation. Dans certains systèmes, les requêtes peuvent être formulées par des phrases. Celles-ci ne sont pas comprises par le système, mais décomposées afin d'en retenir les termes importants.
- il existe également des SRI qui utilisent une approche différente de l'interrogation, principalement basée sur la visualisation globale de l'ensemble des documents, et sur des outils permettant une exploration de cet ensemble.

2.2.1.2 Visualisation de l'information

La restitution d'une liste de documents

Les documents retrouvés par le système à la suite des requêtes peuvent être retournés aux utilisateurs sous des formes différentes. Il est évident que la présentation de chaque document dans son intégralité n'est envisageable que lorsque ceux-ci sont suffisamment courts. De manière générale, on gagne en lisibilité si seuls les titres sont présentés.

Il se trouve que les titres seuls sont souvent insuffisants pour décrire précisément les documents. Il est alors possible d'en afficher les premières phrases, en supposant qu'elles constituent une bonne description du contenu des documents. C'est ce que proposent la plupart des moteurs de recherche sur Internet.

Une technique plus évoluée consiste à remplacer les premières phrases par la partie (ou les parties, suivant leur taille) du document qui contient les termes présents dans la requête. Cette technique est employée, par exemple, par le moteur de recherche dans les groupes de discussion (mailing-lists) ListQuest[1].

Plus évoluée encore, la technique présentée par Tombros et Sanderson (1998) qui proposent un système qui résume automatiquement les documents restitués. Ces résumés sont construits en donnant une importance plus élevée aux phrases qui contiennent des termes présents dans les requêtes, afin qu'ils soient adaptés aux sujets des recherches de chaque utilisateur.

Ainsi, la présentation d'une *partie* ou, mieux encore, du *résumé* des documents restitués en plus de leur titre, permet aux utilisateurs de repérer rapidement les documents pertinents, en montrant dans quel contexte sont utilisés les termes ambigus présents dans les requêtes. Cette technique évite donc aux utilisateurs d'accéder aux documents complets en donnant un aperçu de leur contenu.

La plupart des SRI permettent de retrouver les documents qui correspondent plus ou moins bien aux requêtes. Les documents sont alors retournés par ordre de pertinence (voir 2.2.3.1) décroissante, ceux qui se rapprochent le plus de la requête étant en début de liste. Une *note de pertinence* est parfois indiquée, qui peut être utile lorsqu'une longue liste de documents est retournée. Elle permet de ne s'intéresser qu'aux documents dont la "note" est supérieure à un certain seuil. C'est de cette manière que les résultats de la plupart des moteurs de recherche sur Internet sont présentés. Une présentation similaire est utilisée par

[1] http ://www.listquest.com/

le système documentaire de la NASA, spécialisé dans les articles en astrophysique : ADS[2] abstract service (Kurtz et al., 1993; Accomazzi et al., 1997).

2.2.1.3 Les visualisations graphiques

De nombreux SRI proposent des représentations graphiques de l'information. Nous pouvons en distinguer deux types :

- les représentations **individuelles** où chaque entité (les documents pour les SRI documentaires) est représentée séparément ;
- les représentations **globales** où seuls figurent les caractères généraux des données (les thèmes ou concepts pour les SRI documentaires).

Les représentations graphiques individuelles

Voici quelques-unes des représentations individuelles qui ont été mises au point :

- Le système AIR (Belew, 1989) propose une visualisation des réponses aux requêtes sous la forme d'un schéma bidimensionnel. Sur ce schéma figurent sous la forme d'icônes à la fois les documents, les termes et les liens qui existent entre eux. Les utilisateurs peuvent ainsi connaître les termes qui ont contribué à la restitution des documents, et peuvent affiner leur requête en conséquence.
- Le système TETRALOGIE[3] peut également être utilisé pour la visualisation de documents (Mothe et Dkaki, 1998). Ceux-ci sont placés dans un repère qui définit trois axes correspondant à trois dimensions. L'utilisation de différentes couleurs permet d'y ajouter une quatrième dimension. Cette représentation met en évidence les liens qui peuvent exister entre le contenu des documents, les auteurs, les dates de publications, etc.
- Un système de visualisation des liens entre les documents WEB est proposé par Snowdon : WWW3D (Snowdon et al., 1996). Les documents y sont représentés par des sphères situées dans un espace tridimensionnel. Les liens vers d'autres documents (les ancres) que contiennent les documents sont prises en compte et sont représentés par des segments qui relient les sphères entre elles.

Cette liste de travaux sur la visualisation individuelle des documents n'est bien sûr pas exhaustive. On peut toutefois mettre en avant une limitation propre aux visualisations individuelles : elles deviennent très compliquées et quasiment inutilisables lorsqu'on désire visualiser un grand nombre de documents.

Les représentations graphiques globales

Les représentations de ce type sont généralement issues d'une classification des documents. Ainsi, seuls les représentants des classes de documents figurent sur le graphique. Il est alors possible de visualiser de gros ensembles de documents pour lesquels une représentation individuelle serait inappropriée.

Le système NEURODOC (Lelu et François, 1992) présente les groupes de documents sur un plan (ou une carte) où chaque groupe est décrit par des termes représentatifs des

[2] Astrophysics Data System. http ://ads.harvard.edu, http ://cdsads.u-strasbg.fr

[3] Outil développé à l'Institut de Recherche en Informatique de Toulouse (IRIT) pour la fouille de données (data mining).

documents qui s'y trouvent ainsi que par le nombre de documents qu'il contient. On peut ensuite accéder aux documents en sélectionnant le groupe désiré.

Une représentation assez proche est utilisée dans le ET-SPACE (Chen et al., 1998) : dans ce cas, les groupes de documents sont eux-mêmes situés dans des zones relatives à un thème commun. Ce système présente donc une carte où figurent des zones de différentes couleurs, ainsi que quelques termes décrivant les thèmes abordés dans les documents qui s'y trouvent. Les expérimentations ont porté sur la recherche de sites Web.

Le système WEBSOM (Kohonen et al., 1996) propose une visualisation utilisant les couleurs comme une indication de la densité des documents. Ces derniers sont également répartis sur une carte sur laquelle figurent des termes descriptifs des sujets abordés. De plus, la carte est interactive et offre des possibilités de déplacement et de visualisation rapprochée (zooming). Ce système est clairement adapté à la visualisation d'une grande quantité de documents, les auteurs proposant une visualisation de plus d'un million de documents issus de quatre-vingts groupes de discussion.

La plupart des systèmes qui proposent une représentation globale des documents sont basés sur un réseau de neurones particulier : une carte auto-organisatrice (Kohonen, 1982) que nous décrirons dans le chapitre suivant. Ce type de réseau a la capacité de fournir une répartition bidimensionnelle des données qu'on lui applique. Cette répartition est telle que les données proches sur la carte sont de caractéristiques voisines. Ce genre de cartes est à l'origine, par exemple, de la présence des zones colorées où se situent les documents traitant globalement d'un même thème dans la visualisation du ET-SPACE. D'une manière générale, les cartes auto-organisatrices sont bien adaptées à la visualisation d'un ensemble de données (Vesanto, 1999).

2.2.2 Organisation interne des données

Un SRI doit pouvoir accéder à l'information contenue dans les documents pour effectuer les recherches. Ces données pourront être organisées de différentes façons en fonction notamment de la taille de la base textuelle.

2.2.2.1 La recherche dans les textes bruts

La méthode la plus simple consiste à effectuer les recherches directement dans les textes des documents pour déterminer quels sont les documents qui contiennent les termes en rapport avec les requêtes. Pour cela, différents algorithmes ont été mis au point en vue de diminuer la durée des recherches, qui est fonction de deux paramètres :

- la longueur du terme à rechercher ;
- la longueur du texte à analyser ;

Le meilleur algorithme (Boyer-Moore-Horspool) fonctionne environ 8 fois plus rapidement que le moins bon (brute force algorithm). Une description des principaux algorithmes de recherche de chaînes de caractères se trouve dans un article de Baesa-Yates (1992).

Ces méthodes de recherche, si elles sont efficaces, sont toutefois lentes car les textes des documents comportent un grand nombre de termes *vides*[4] qui allongent inutilement la taille du texte où s'effectuent les recherches. Pour réduire les temps de recherche, on a alors recours aux fichiers inverses (voir 2.2.2.3).

[4]Les termes vides sont les termes qui n'ont pas de signification : les articles par exemple.

Terme	Doc. 1	Doc. 2	Doc. 3	Doc. 4	Doc. 5
étoiles	X				X
galaxies	X		X		
asteroïdes		X		X	X
...	...	...	...	...	...

Tab. 2.1: *Table reliant les termes aux documents qui les contiennent*

2.2.2.2 L'indexation

Pour remédier aux problèmes causés par la place importante que prennent les données en mémoire, ainsi que le ralentissement des recherches dû au grand nombre de termes *vides* dans les textes, on a généralement recours à l'*indexation* (à laquelle nous allons revenir en détail dans la section 2.3). L'indexation consiste simplement à choisir (de manière automatique ou manuelle) quelques termes (ou mots-clés) pour seuls descriptifs des documents. Ainsi, un document traitant du froid dans le grand nord pourra être décrit par les termes "pôle nord", "température", "conditions de vie" et "météorologie".

Nous voyons bien l'avantage, le volume des données dans lesquelles les recherches sont effectuées se trouve considérablement réduit. De plus, les termes inutiles ou ambigus, lorsqu'ils sont sortis de leur contexte, ont été éliminés, ce qui évite la restitution par les systèmes de documents inadéquats. En revanche, le choix des termes d'indexation est *primordial* car un document défini par de mauvais mots-clés a toutes les chances de ne pas être retrouvé alors qu'il devrait l'être.

2.2.2.3 Les fichiers inverses

Un fichier inverse est une table de correspondance qui associe des termes aux documents qui les contiennent (Tab. 2.1).

Grâce à cette table, il est possible de savoir quels documents contiennent tel terme sans avoir à parcourir l'ensemble de la base textuelle, mais uniquement en parcourant les entrées du fichier inverse pour y trouver le terme recherché. En fait, cela revient à effectuer une recherche préalable avec différents termes, puis à stocker les résultats des recherches, afin qu'ils soient réutilisables. De ce fait, la création d'un fichier inverse nécessite un traitement[5] préliminaire qui consiste à :

- repérer tous les termes présents dans les textes de la base, ou ceux qui sont issus d'une indexation ;
- mémoriser les documents dans lesquels ils se trouvent.

L'utilisation d'un fichier inverse permet des recherches très rapides (plusieurs ordres de grandeurs par rapport aux recherches traditionnelles), et est donc conseillée pour traiter les grandes bases de données. L'inconvénient est qu'un fichier inverse peut nécessiter beaucoup de place, et peut parfois être plus volumineux que les données elles-mêmes (Haskin, 1981). D'autre part, la mise à jour des fichiers inverses est assez coûteuse en ressources informatiques et peut devenir gênante si elle doit être effectuée fréquemment.

On pourra trouver une bonne description des fichiers inverses dans Harman et al. (1992), où différentes implémentations sont suggérées.

[5] Cette opération se fait bien entendu de manière totalement automatique.

2.2.3 Évaluation des systèmes de recherche d'information

Il est possible d'évaluer les SRI sur différents points :

- la présentation des résultats ;
- les temps d'attente entre la formulation d'une requête et la restitution des documents correspondants ;
- l'effort demandé aux utilisateurs pour obtenir une réponse qui leur convient ;
- la qualité des résultats des recherches.

C'est sur ce dernier point que nous allons nous attarder ici. Pour quantifier la qualité des résultats de recherche d'un SRI, deux mesures sont communément employées : les *taux de rappel* et *taux de précision*. Mais auparavant, il est important de définir ce que l'on entend par la *pertinence* d'un document.

2.2.3.1 La pertinence des documents

La pertinence d'un document caractérise son adéquation avec une requête, c'est à dire la validité de l'information qu'il apporte en réponse à la question que l'utilisateur se pose. Nous parlons ici de la *pertinence utilisateur* qu'il convient de différencier de la *pertinence système* qui est évaluée par les SRI (voir le chapitre 4). La pertinence utilisateur est donc une notion très subjective car plusieurs personnes peuvent avoir des avis différents sur le bon accord entre un document et une requête donnée. Fort heureusement, il semble que les différences soient suffisamment faibles pour que des études faisant intervenir le jugement de personnes sur la pertinence des documents retrouvés par les SRI aient un sens (Van Rijsbergen, 1979). Néanmoins, des études montrent que les résultats d'évaluations de SRI utilisant des sujets auxquels correspondent peu de documents pertinents (autour de 5) ne sont pas exploitables, car celles-ci ne donnent pas des résultats suffisamment cohérents d'une personne à l'autre (Voorhees, 1998).

2.2.3.2 Taux de précision, taux de rappel

La notion de pertinence étant posée, nous pouvons définir les taux de précision TP et taux de rappel TR :

$$TP = \frac{|P \cap R|}{|R|}$$

$$TR = \frac{|P \cap R|}{|P|}$$

où, **pour une requête donnée** : P est l'ensemble des documents pertinents contenus dans la base et R est l'ensemble des documents restitués comme résultat à la requête.

Nous voyons que pour un ensemble de documents restitués, un couple de valeurs (TP, TR) peut être calculé. Si $TR = 1$ et $TP = 1$, le système est parfait : il n'a renvoyé non seulement que des documents pertinents ($TP = 1$) mais également tous les documents pertinents ($TR = 1$). Évidemment, aucun SRI parfait n'a encore été mis au point, et les valeurs TP et TR varient entre 0 et 1 de manière anticorrélée.

Lorsqu'un SRI a tendance à restituer un grand nombre de documents, la probabilité de retrouver des documents pertinents est assez grande $TR \approx 1$, mais aussi celle de retrouver des documents non pertinents : $TP \approx 0$. Inversement, si un SRI restitue une petite quantité

de documents à la suite d'une interrogation, seuls les documents qui semblent les plus pertinents (selon la technique d'évaluation) s'y trouvent. La probabilité d'observer une majorité de documents non pertinents est donc faible : $TP \approx 1$ et celle d'y trouver tous les documents pertinents est faible également $TR \approx 0$.

Si, pour une requête donnée, il est possible de faire varier un paramètre m influant sur les documents restitués, alors on peut tracer la variation de TP en fonction de TR. Le paramètre m peut être le nombre de documents retournés par exemple. Ainsi, si un SRI renvoie 100 documents ordonnés par ordre de pertinence décroissante, on peut calculer plusieurs couples (TP,TR) en ne tenant compte par exemple que des dix premiers documents retournés, puis des vingt premiers, des trente premiers etc... On peut aussi calculer TP pour des valeurs données de TR. Il est ensuite possible de tracer la moyenne de courbes correspondant à différentes requêtes (Fig. 2.2).

Il est important de noter que l'évaluation des taux de précision et de rappel repose sur la connaissance *a priori* de tous les documents de la base qui sont pertinents pour une requête donnée. Dans les expérimentations où plusieurs SRI sont utilisés, on contourne la difficulté avec la mise en commun des documents pertinents retournés par tous les SRI. On considère alors que ces documents forment la totalité des documents pertinents de la base.

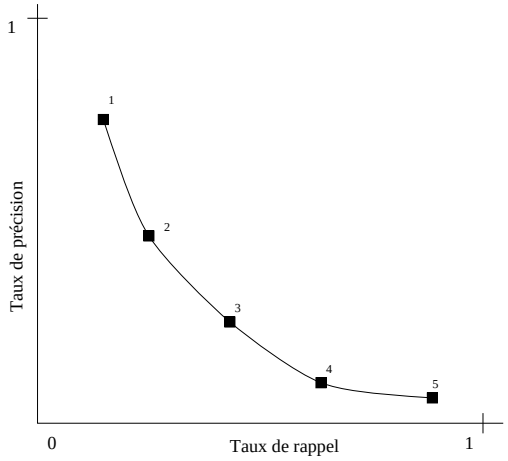

Fig. 2.2: *Courbe rappel-précision pour une requête donnée ou courbe moyennée sur plusieurs requêtes. Chaque point correspond à une valeur d'un paramètre variable.*

Afin de comparer différents SRI entre eux, des bases textuelles ont été mises au point, pour lesquelles des requêtes et les documents pertinents correspondants sont connus (généralement déterminés au préalable par des experts). On peut citer la base de l'université de Cranfield (Cleverdon, 1962) qui compte 1400 documents et environ 400 requêtes et les documents pertinents associés, dans le domaine de l'aéronautique. Plus récemment, la première conférence TREC[6] a été organisée (Harman, 1993), et depuis lors, les participants montrent aux conférences annuelles TREC les résultats de leurs études sur des bases textuelles communes.

Les SRI peuvent être testés non seulement sur les documents qu'ils permettent de retrouver, mais aussi sur leur interaction avec les utilisateurs. Ainsi, des évaluations ont été menées au cours desquelles les utilisateurs devaient retrouver des documents correspondant à des requêtes préétablies en un temps limité de vingt minutes (Lagergren, 1998).

[6]Text REtrieval Conference

2.2.4 Conclusion

Nous avons présenté certains aspects des SRI, en particulier l'interface d'utilisation, l'organisation des données dans lesquelles les recherches sont effectuées et l'évaluation de leurs performances. Nous nous attacherons, dans le chapitre 4, à décrire les méthodes utilisées dans ces systèmes pour évaluer la pertinence des documents aux requêtes des utilisateurs.

2.3 La représentation des documents

L'information sémantique contenue dans les documents est directement reliée aux mots qu'ils contiennent. Il est évident que parmi tous les mots qui constituent un texte, certains sont moins significatifs que d'autres, et certains ne sont souvent même pas significatifs du tout.

Il est donc possible de construire le descripteur de chaque document avec les mots les plus significatifs, c'est à dire qui reflètent bien leur contenu. Cette tâche s'appelle l'indexation.

2.3.1 Les termes d'indexation

L'expression "terme d'indexation" désigne les termes, ou mots-clés, qui sont attribués aux documents d'une base pour les décrire. Un mot-clé peut être constitué d'un mot simple, ou encore d'un groupe de mots.

2.3.2 L'indexation manuelle

L'indexation des documents d'une base de données textuelles peut être confiée à des personnes. Celles-ci doivent lire les textes afin de pouvoir choisir les termes (mots-clés) qui les décrivent le mieux. En fonction du contenu de la base, il est souvent nécessaire que les personnes chargées de cette tâche soient spécialistes des domaines traités dans les documents.

Le travail d'indexation devient moins fastidieux s'il est fait par les auteurs des documents eux-mêmes. C'est de cette manière que les articles publiés dans le journal "Astronomy and Astrophysics" sont indexés : les auteurs doivent choisir dans une liste de mots-clés ceux qui conviennent le mieux à leurs publications.

Le principal problème posé par l'indexation manuelle est sa lenteur (sauf dans le cas où les auteurs des documents font ce travail). Tous les textes doivent être lus, compris, et synthétisés manuellement, ce qui rend impossible l'indexation des textes de grandes bases mises à jour fréquemment.

Un autre problème important est celui du choix des termes d'indexation. Il est évident que deux personnes différentes n'associeront pas les mêmes termes à un même document. De plus, même une personne habituée à indexer des documents dans un domaine donné peut choisir des termes différents pour un même document à des moments différents (Bates, 1986).

2.3.3 L'indexation automatique

L'indexation automatique est assurée par un programme informatique dont l'action consiste à parcourir les textes afin de répertorier et compter tous les termes rencontrés

dans les documents. Les termes significatifs sont ensuite déterminés et les moins significatifs éliminés.

2.3.3.1 Les anti-dictionnaires

Un anti-dictionnaire est une liste de termes qui ne sont porteurs d'aucune information sémantique (ils sont dits termes ou mots *vides*). C'est le cas des prépositions, pronoms, articles, etc... Ces mots se retrouvent dans tous les textes et sont indépendants des thèmes traités.

Les mots contenus dans un anti-dictionnaire sont automatiquement éliminés.

2.3.3.2 La loi de Zipf et l'élimination des termes peu fréquents

La loi de Zipf

La loi empirique de Zipf (1949) relie les fréquences d'apparition des termes dans les textes f_i à leur rang[7] r_i . La figure 2.3 illustre cette loi qui stipule que :

$$f_i.r_i \approx K$$

où K est une constante. Cette loi s'applique aussi bien aux **fréquences relatives** des termes (leur nombre d'occurence dans un même document) qu'à leur **fréquence absolue** (le nombre de documents dans lesquels ils apparaissent). Elle montre qu'une grande quantité de termes rencontrés dans les textes sont peu fréquents. Typiquement, la moitié des termes n'apparaissent qu'une seule fois.

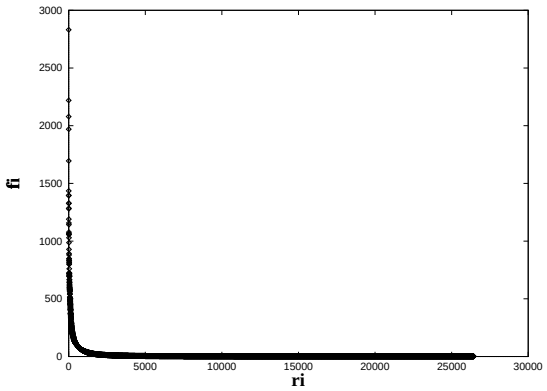

Fig. 2.3: *La fréquence absolue f_i des termes en fonction de leur rang r_i. Ces termes, au nombre de 26444, sont tirés des enregistrements de 12553 articles parus dans la revue "Astrophysical Journal" (termes tirés des titres, résumés et mots-clés, les termes vides ont été éliminés auparavant). 59% de ces termes apparaissent dans un seul document.*

L'élimination des termes peu fréquents

La loi de Zipf montre que les termes peu fréquents sont les plus nombreux. Le nombre des termes d'indexation peut, suivant la taille de la base textuelle, être très important (voir l'exemple de la figure 2.3) et il est courant de réduire ce nombre en éliminant les termes

[7]Le terme le plus fréquent est de rang 1 tandis que le terme le moins fréquent est de rang maximal (variable suivant le corpus).

les moins fréquents. Deux justification à cette élimination (ou coupure) existent, suivant qu'elle est basée sur les valeurs de la fréquence relative ou absolue des termes.

- **la coupure basée sur les fréquences relatives :** les termes peu fréquents dans un document sont considérés comme peu représentatifs de son contenu et ne sont donc pas une source importante d'information pour la description du document en question ;
- **la coupure basée sur les fréquences absolues :** cette coupure est fréquemment utilisée lorsqu'une classification des documents d'une base est envisagée. Classer des documents consiste à regrouper les documents qui comportent les mêmes termes. La présence de termes qui apparaissent dans peu de documents est néfaste car elle tend à créer de petits groupes et à éclater et disperser les groupes représentatifs des grands thèmes traités dans le corpus. Les termes de faible fréquence absolue sont parfois considérés comme du bruit (Chen et al., 1998).

Remarque : Il n'existe pas de moyens théoriques pour déterminer la fréquence (relative ou absolue) en deça de laquelle les termes sont éliminés, ceci doit se faire de manière empirique (Belew, 1999).

2.3.3.3 La recherche des radicaux

Cette méthode consiste à ne prendre en considération que les racines des termes rencontrés dans les textes. De cette façon, on réduit le nombre total de termes d'indexation, puisque tous les mots ayant une même racine sont fusionnés en un même mot-clé. Par exemple, les termes "computer", "compute" et "computing" sont assimilés à l'unique terme "comput".

L'algorithme le plus couramment utilisé pour la recherche des radicaux de mots de langue anglaise est celui de Porter (Porter, 1980), mais il n'est pas infaillible : par exemple les termes "corona" et "coronal" sont remplacés respectivement par "corona" et "coron". Notons que l'équivalent de l'algorithme de Porter pour la langue française n'existe pas, et la meilleure méthode dans ce cas consiste à tronquer les mots pour ne conserver que les premiers caractères.

2.3.3.4 La pondération des mots-clés

Les méthodes de pondération des mots-clés ont pour but de donner plus de poids (plus d'influence lors des recherches relatives aux requêtes) à certains termes jugés plus importants que d'autres. L'importance des termes est déduite automatiquement en fonction de leurs fréquences d'apparition dans les textes. Nous décrivons ici la pondération des termes la plus courante, celle que Sparck Jones proposa en 1972 et que nous utiliserons dans notre système. Le calcul des poids des termes suit les deux tendances ci-dessous (Sparck Jones, 1972) :

- plus le nombre de documents qui contiennent un terme i est grand, moins ce terme doit avoir un poids fort. On comprend aisément qu'un terme qui apparaît dans la plupart des documents a un pouvoir discriminant moindre par rapport à un terme qui apparaît dans peu de documents qui, lui, est plus spécialisé.

  On fait pour cela appel à la notion de fréquence absolue inverse idf_i du document i qui s'écrit :

Terme	Doc. 1 binaire	Doc. 2 binaire	Doc. 1 numérique	Doc. 2 numérique
étoiles	1	0	0.5	0
galaxies	1	0	0.1	0
asteroïdes	0	1	0	0.9
...	...	...	...	...

Tab. 2.2: *Vecteurs binaires et numériques pour deux documents.*

$$idf_i = Log(\frac{N}{f_i})$$

où N représente le nombre de documents de la collection, et f_i la fréquence absolue du terme i (le nombre de documents qui contiennent le terme i).

– plus un terme est fréquent dans un document donné (fréquence relative), plus il est caractéristique de ce document. Son poids doit donc être supérieur à celui d'un terme moins fréquent dans le même document. Le poids du terme i dans un document d doit suivre les mêmes variations que sa fréquence relative tf_{di}.

Les deux tendances précédentes sont combinées de la manière suivante :

$$w_{di} = tf_{di}.idf_i$$

où w_{di} est le poids du terme i dans le document d.

2.3.4 Les vecteurs documents

Après l'indexation des documents (indexation manuelle ou automatique) on dispose d'un certain nombre de termes qui définissent à eux seuls les documents de la base. Afin de simplifier la manière de traiter les document par les SRI, il est commode d'associer un vecteur à chaque document, vecteur dont chacune des composantes est relative à l'un des termes d'indexation. Les vecteurs sont donc de même dimension : le nombre de composantes étant égal au nombre de termes d'indexation pour l'ensemble de la collection.

2.3.4.1 Les vecteurs binaires

Ces vecteurs sont utilisés dans les petites bases où une pondération des mots-clés n'aurait pas un effet important, et dans le cas où l'indexation est faite manuellement et où aucune pondération n'est utilisée.

Les vecteurs sont dans ce cas constitués de 0 et de 1, une composante nulle signifiant que le terme correspondant est absent du document, et une composante non-nulle signifiant que le terme correspondant est présent dans le document (voir le tableau 2.2).

2.3.4.2 Les vecteurs numériques

Ces vecteurs suivent le même principe que les vecteurs binaires décrits ci-dessus, avec pour différence que les 1 sont remplacés par des valeurs numériques variables : les poids w_{di} calculés en fonction des fréquences d'apparition des mots-clés (voir 2.3.3.4 ainsi que le tableau 2.2).

2.3.5 Conclusion

2.3.5.1 L'indexation

L'indexation des documents, qui consiste à leur attribuer des termes en tant que descripteurs, permet d'accélérer les recherches, puisque celles-ci ne sont plus effectuées sur l'ensemble des textes des documents, mais sur les termes d'indexation uniquement. Cet avantage est suivi d'un danger : puisque seuls les termes d'indexation sont utilisés pour faire les recherches, il faut absolument que ces termes soient bien choisis.

2.3.5.2 La conservation de l'information

Il est important de noter que :

- le choix des termes d'indexation est primordial pour la recherche d'information dans les documents indexés. Des termes mal choisis pourront donner des documents inappropriés en réponse à des requêtes, mais aussi empêcher la restitution de documents pertinents. De manière générale, les termes d'indexation associés à un document sont plus nombreux lorsqu'ils sont issus d'une méthode automatique d'indexation plutôt que dans le cas d'une indexation manuelle.
- toute l'information sémantique n'est pas conservée. Dans le cas optimal où les termes d'indexation sont tous bien choisis sans qu'aucun ne soit oublié, l'information sur les positions relatives des termes est quand-même perdue. Néanmoins, même si la représentation des documents après l'indexation ne permet pas de renseigner précisément sur le sens des textes, elle renseigne encore fort bien sur les thèmes qui y sont traités, et c'est ce qui importe en recherche d'information.