Chapitre 4

Les Systèmes de Recherche d'Information

Ce chapitre a pour objet de décrire les types de systèmes de recherche d'information (SRI) les plus fréquemment utilisés. Il est possible de classer les différents SRI dans quatre groupes, suivant les méthodes sur lesquelles ils sont basés afin de retrouver les documents pertinents à une requête :

- certains adoptent une représentation vectorielle des requêtes et des documents, puis procèdent à des calculs de ressemblance ;
- d'autres effectuent des opérations booléennes sur les documents ;
- d'autres utilisent les réseaux de neurones ;
- d'autres encore font appel à des calculs de probabilité.

Différentes techniques peuvent ensuite se "greffer" sur ces modèles et les compléter, telles que la reformulation des requêtes ou encore les classifications, qui visent chacune à leur manière à améliorer les résultats de recherche.

4.1 L'approche vectorielle

Le modèle vectoriel a été proposé dès les années 70 par SALTON avec le système SMART (Salton, 1971). Il repose sur :

- une *représentation similaire* des documents et des requêtes par des vecteurs de même type ;
- l'hypothèse que les documents les plus pertinents sont ceux qui sont les plus proches des requêtes (qui contiennent les mêmes termes).

Puisque la même représentation est utilisée pour les documents contenus dans la base et pour les requêtes des utilisateurs, il est possible de comparer directement les documents aux requêtes, et leur attribuer un *degré de ressemblance*. Ainsi, la *pertinence système* des documents s'en déduit suivant la règle : *plus un document est proche d'une requête* (plus il lui ressemble), *plus il est pertinent*. L'utilisateur peut donc retrouver des documents plus ou moins pertinents, qui lui sont d'ailleurs généralement retournés par ordre de pertinence (système) décroissante pour plus de clarté. On échappe ainsi à la limitation du modèle booléen qui permet de retourner uniquement des documents qui correspondent exactement aux requêtes.

Les vecteurs utilisés sont soit binaires, indiquant simplement la présence ou l'absence des termes dans le document ou la requête, soit numériques où une pondération des termes est possible (voir "La pondération des mots-clés" dans la section 2.3.3.4).

La mesure de similarité

L'attribution d'une note de pertinence (système) aux documents, ce qui revient à faire une comparaison de vecteurs, peut se faire de différentes manières : par exemple par un produit scalaire, par la mesure de l'angle que forment les vecteurs, ou encore par une mesure de distance.

Notons $\mathbf{D}$ un vecteur document, et $\mathbf{R}$ un vecteur requête. La mesure de similarité $SIM(\mathbf{D},\mathbf{R})$ peut être calculée selon :

$$SIM(\mathbf{D},\mathbf{R}) = \sum_{i=1}^{n} D_i.R_i \qquad \text{(produit scalaire)}$$

ou :

$$SIM(\mathbf{D},\mathbf{R}) = \frac{\sum_{i=1}^{n} D_i.R_i}{\sqrt{\sum_{i=1}^{n} D_i^2 . \sum_{i=1}^{n} R_i^2}} \qquad \text{(cosinus)}$$

on peut aussi utiliser une mesure de distance :

$$DIST(\mathbf{D},\mathbf{R}) = \sqrt{\sum_{i=1}^{n} (D_i - R_i)^2} \qquad \text{(distance euclidienne)}$$

On peut noter que les deux premières formulations donnent des valeurs grandes pour des vecteurs D et R proches, et nulles pour des vecteurs totalement différents ; on peut alors assimiler la pertinence système d'un document à l'une ou l'autre de ces mesures :

$$Pert(\mathbf{D},\mathbf{R}) \equiv SIM(\mathbf{D},\mathbf{R})$$

Cependant, la distance entre les vecteurs varie dans le sens opposé de la pertinence et on pourra par exemple utiliser son inverse :

$$Pert(\mathbf{D},\mathbf{R}) \equiv 1/DIST(\mathbf{D},\mathbf{R})$$

Remarque : D'autres mesures de similarité sont utilisables, se situant à mi-chemin entre le modèle vectoriel et le modèle booléen. Leurs caractéristiques sont les suivantes :

- elles ne font pas appel à des opérations vectorielles (calculs de produit scalaire, d'angle ou de distance) ;
- elles rendent possible un degré de pertinence système ;
- elles peuvent utiliser des vecteurs binaires.

En voici quelques-unes :

cœfficient de Dice : $2.\frac{|D \cap R|}{|D| + |R|}$

cœfficient de Jaccard : $\frac{|D \cap R|}{|D \cup R|}$

mesure de recouvrement : $\frac{|D \cap R|}{min(|D|, |R|)}$

4.2 L'approche booléenne

En ce qui concerne la représentation des documents, le modèle de recherche booléen peut être considéré comme la plus simple des représentations vectorielles (voir 2.3.4). Celle où les composantes des vecteurs sont *binaires*, c'est-à-dire qu'elles indiquent uniquement la présence ou l'absence des termes d'indexation dans les documents.

Toutefois, les modèles booléen et vectoriel sont très différents sur un point : la représentation des requêtes. Avec le modèle booléen, une requête correspond à une série d'opérations logiques qui permettent d'éliminer ou de conserver les documents en fonction de leur contenu. Par exemple, il est possible de retrouver les documents qui contiennent les termes A et C, sans le terme B (fig. 4.1). Il est alors évident que :

- les requêtes doivent être formulées avec les termes *exacts* d'indexation (combinés avec les opérateurs logiques OU, ET et NON) ;
- les documents retournés sont ceux qui correspondent *exactement* aux requêtes ;
- les réponses du système *ne peuvent pas* être ordonnées de manière à mettre en avant les documents susceptibles d'intéresser davantage les utilisateurs ;

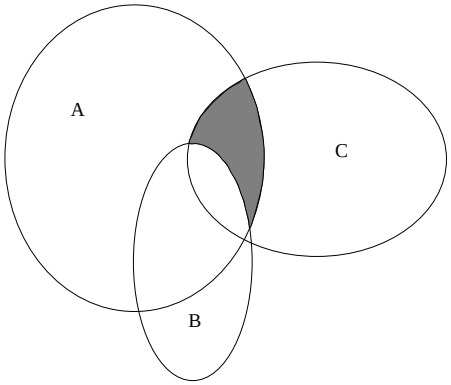

Fig. 4.1: *Documents correspondant à la requête : A et C et NON B.*

La description complète, ainsi que les détails techniques d'une implémentation d'un système booléen est proposée par S. Wartik (Wartik, 1992). Néanmoins, il semble que le modèle booléen soit de moins en moins utilisé, au profit du modèle vectoriel.

4.3 L'approche neuronale

L'idée d'utiliser les réseaux de neurones dans le domaine de la recherche d'information est apparue dans les années quatre-vingt, avec les travaux de M. Mozer (Mozer, 1984).

Nous classerons les SRI basés sur les réseaux de neurones en deux familles suivant le type du réseau utilisé :

- ceux qui font appel à un apprentissage supervisé ;
- ceux qui font appel à un apprentissage non supervisé.

Les systèmes qui en résultent sont assez différents. Les premiers utilisent les requêtes comme des stimuli qui se propagent dans le réseau jusqu'aux neurones de sortie. Pour les seconds, le réseau est utilisé pour effectuer une classification des documents.

4.3.1 Les systèmes à apprentissage supervisé

L'architecture

De manière générale, ces systèmes sont construits autour d'un réseau à couches. Ainsi, on peut avoir une couche dédiée aux requêtes, une autre aux termes d'indexation, et une autre encore aux documents (Kwok, 1989). D'autres systèmes sont quant à eux constitués de deux couches : une couche "documents" et une couche "termes d'indexation"[1] (Wilkinson et Hingston, 1991; Belew, 1989). Finalement, chaque neurone correspond à une entité : un document ou un terme. Ainsi, en général, les termes d'une requête excitent les neurones correspondants de la couche des requêtes (ou à défaut, ceux de la couche des termes), l'activation se propage vers la couche de sortie (couche "documents") et la pertinence système des documents se déduit alors de l'activation des neurones correspondants.

L'apprentissage

Les réseaux qui composent ces systèmes ne nécessitent pas d'apprentissage afin d'être opérationnels. Ils sont initialisés en fonction des connaissances dont on dispose sur les données : les fréquences relative et absolue des termes dans les documents, la taille des documents ou encore la co-occurrence de termes etc. Certains systèmes sont initialisés avec les pondérations issues de la représentation vectorielle. Ils donnent alors des résultats similaires à ceux d'un système vectoriel lorsque le réseau qui les constituent n'a pas subit d'apprentissage (Wilkinson et Hingston, 1991; Mothe, 1994).

L'apprentissage intervient ensuite dans le but d'affiner les résultats. Par exemple, les poids du réseau peuvent être modifiés en fonction du jugement des utilisateurs sur la pertinence[2] des documents retournés. Ces poids peuvent être ceux des liens qui relient les neurones "documents" aux neurones "termes" correspondant à la requête, mais aussi, lorsqu'ils existent, des liens entre les neurones termes (Wong et al., 1993; Mothe, 1994).

4.3.2 Les systèmes à apprentissage non supervisé

Ils sont basés sur des réseaux de type "winner takes all" (WTA) ou "winner takes most" (WTM) que nous avons décrits en 3.4. L'apprentissage effectue une classification des documents, où chaque classe est représentée par un neurone de sortie. Les entrées du réseau, utilisées durant l'apprentissage sont donc constituées par des vecteurs descriptifs des documents. Par exemple, dans le système NEURODOC (Lelu et François, 1992), on effectue une classification des documents en utilisant un apprentissage inspiré de l'algorithme des K-Means.

La plupart des systèmes basés sur un réseau de neurones à apprentissage non supervisé sont de type SOM. Lin fut l'un des premiers à mettre au point un SRI utilisant ces réseaux. Les expérimentations ont porté sur une collection de 140 documents (Lin et al., 1991).

Les SOM sont maintenant beaucoup employées dans des systèmes dédiés à la recherche sur Internet : on peut citer les systèmes SiteMaps (Lin, 1997) et ET-SPACE (Chen et al., 1996) dédiés à la recherche de pages WEB, ainsi que le système WEBSOM de l'équipe de Kohonen (Kohonen et al., 1996) qui s'est orientée vers la cartographie de messages parus dans les groupes de discussion. Néanmoins, ces systèmes sont encore à l'état de prototypes et les moteurs de recherche courants ne les ont pas encore adoptés.

[1] Certains systèmes (Belew) utilisent aussi les noms des auteurs.

[2] Il est ici question de pertinence utilisateur.

4.4 L'approche probabiliste

Les modèles probabilistes sont très différents de ceux que nous venons de passer en revue, car ici, ce n'est pas un degré de pertinence (système) qui est calculé pour la recherche des documents, ni une adéquation des documents avec des requêtes booléennes. L'approche probabiliste s'intéresse à la *probabilité de pertinence* des documents (Maron, 1965) (Robertson et Sparck Jones, 1976). Ainsi, dans cette approche, les documents ne sont pas plus ou moins pertinents, mais leur probabilité de pertinence est plus ou moins importante.

La théorie des probabilités forme donc la base de ces modèles, et tout particulièrement le théorème de Bayes[3],

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)}$$

grâce auquel la probabilité de pertinence peut être évaluée pour chaque document (Robertson et Sparck Jones, 1976). Ce calcul des probabilités de pertinence repose sur :

- une indépendance des termes d'indexation
- une évaluation du nombre d'occurrences des termes d'indexation dans les documents pertinents et non-pertinents ; ceci peut se faire en utilisant une distribution a priori des termes (loi de Poisson) ou par une étude préliminaire d'un échantillon de documents.

L'indépendance des termes d'indexation est toutefois problématique : en effet, le terme "carcéral" par exemple a certainement une probabilité plus grande d'apparaître dans un texte qui contient "détenus" que dans un autre contenant le terme "carburateur". Ainsi, dès 1964, Williams (Williams, 1965) indique que l'indépendance des termes d'un document est supposée pour des raisons purement mathématiques, sans quoi la plupart des calculs ne pourraient être effectués.

4.5 Les reformulations des requêtes

Le but commun à ces techniques est principalement l'amélioration de la qualité des résultats de recherche lorsque la seule évaluation de la similarité entre les requêtes et les documents n'est plus suffisante. C'est le cas par exemple lorsqu'une requête exprime un concept présent dans la base documentaire, et où ce concept est exprimé dans les documents avec des termes différents. Pour pallier ces problèmes, des mécanismes de reformulation des requêtes ont été mis au point. Ils consistent à modifier les requêtes en vue d'améliorer les résultats de la recherche.

Ces techniques sont indépendantes du type de système de recherche : elles peuvent être utilisées aussi bien dans un modèle vectoriel, probabiliste ou neuronal.

4.5.1 La ré-injection de la pertinence

Le principe général de la ré-injection de la pertinence[4] se résume ainsi :

- les utilisateurs effectuent une première requête ;

[3] où $P(A)$ est la probabilité d'observer l'évènement A et $P(A|B)$ la probabilité d'observer l'évènement A sachant que B a été observé.

[4] Il est question ici de pertinence utilisateur.

- des documents sont retournés en fonction de cette première interrogation ;
- les utilisateurs doivent ensuite indiquer parmi les documents retournés, quels sont ceux qui sont pertinents, et/ou ceux qui ne le sont pas ;
- la requête de départ est alors modifiée automatiquement pour tenir compte du jugement des utilisateurs.

En fait, il s'agit de simplifier la tâche aux utilisateurs, qui ne sont plus obligés de rechercher les termes importants dans les articles pertinents, puis d'effectuer une nouvelle requête ; c'est le système qui le fait pour eux. La reformulations de la requête peut se faire par une re-pondération des termes (Sparck Jones, 1979) ou par l'ajout (ou le retrait) de termes contenus dans les documents pertinents à la requête (non pertinents) (Salton et Buckley, 1990). Dans le système de Kwok (Kwok, 1989) qui fait appel à un réseau de neurones, une reformulation de requête est effectuée par un apprentissage qui renforce les liens entre les neurones de requête et ceux des termes.

De manière générale, la requête est modifiée pour ressembler d'avantage aux documents jugés pertinents, et s'éloigner des documents non pertinents.

4.5.2 L'expansion des requêtes

Comme pour la ré-injection de la pertinence, l'expansion des requêtes consiste à ajouter d'autres termes à ceux choisis par les utilisateurs pour l'interrogation. La différence est que ces termes sont choisis automatiquement, sans l'intervention des utilisateurs. Les nouveaux termes doivent avoir des sens proches des termes de ceux donnés par l'utilisateur. Ceci a pour effet de diversifier les documents pouvant répondre à l'interrogation, tout en restant dans le même domaine. Les termes ajoutés peuvent être déduits d'un thésaurus ou d'une étude statistique de l'apparition des termes dans les documents de la base (co-occurrences).

Les systèmes basés sur les réseaux de neurones peuvent également effectuer des expansions de requêtes. Par exemple, dans le système de Wilkinson (Wilkinson et Hingston, 1991), les stimuli peuvent se déplacer dans deux directions : de la couche "termes" à la couche "documents" lorsqu'une requête est posée, et de la couche "documents" à la couche "termes" pour effectuer une expansion de requête. Dans ce cas, l'activation des neurones "documents" qui correspondent aux documents retournés par le système excite les neurones "termes" des termes contenus par ces documents, et une nouvelle activation des neurones "documents" en est déduite. Finalement, cette opération consiste à ajouter les termes que contiennent les documents "jugés" pertinents par le système à la requête de départ (pertinence système).

Les techniques de reformulation des requêtes (expansion automatique ou ré-injection de la pertinence) peuvent être utilisées dans tous les types de SRI ; elles peuvent aussi être combinées entre elles. Néanmoins, une étude montre que l'ajout d'une expansion de requêtes (automatique) à une ré-injection de pertinence (suivant le jugement des utilisateurs) n'améliore pas la qualité des résultats obtenus avec la ré-injection de pertinence seule (Boughanem et al., 1999).

Les améliorations apportées par les reformulations des requêtes sont variables d'une base documentaire à l'autre, et peuvent dépendre par exemple du nombre de termes ajoutés (Qiu et Frei, 1993; Harman, 1992) et, bien sûr, de la pertinence du choix de ces termes.

L'utilisation de ces techniques peut poser des problèmes : ainsi, la ré-injection de la pertinence est d'un emploi souvent lourd pour les utilisateurs qui doivent dialoguer avec

le système, tandis que les termes ajoutés lors d'expansions automatiques des requêtes ne sont pas toujours appropriés.

4.6 La classification

La classification des documents consiste à créer des catégories dans lesquelles les documents sont rangés. Ces catégories peuvent être créées soit manuellement par un expert qui connait la base, soit déduites automatiquement en fonction du contenu de la base à l'aide d'un algorithme approprié. Ensuite, chaque document est classé dans la catégorie dont il est le plus proche. Pour cela, les documents sont comparés au représentant de chaque classe, le **centroïde**, dont les caractéristiques sont généralement la moyenne des caractéristiques des documents appartenant aux classes.

L'intérêt majeur de la classification est donc de regrouper les documents similaires, de manière à faciliter à un utilisateur, à partir d'un document pertinent, l'accès aux documents qui peuvent lui convenir (car proches d'un document pertinent). L'hypothèse de base ici est que *les documents associés dans une même classe ont tendance à être pertinents pour les mêmes requêtes.* On peut trouver une discussion sur cette hypothèse ("the cluster hypothesis") dans (Van Rijsbergen, 1979).

Nous allons maintenant passer en revue quelques méthodes de classification fréquemment rencontrées en recherche d'information.

4.6.1 Classifications non-hiérarchiques

Il s'agit ici de répartir les N documents d'une base dans M classes (avec $N >> M$). Chaque classe est représentée par un *attracteur*, qui peut être un document de la base, ou un document type. Cet attracteur est utilisé pour attirer dans la classe qu'il représente les documents qui lui ressemblent. On peut utiliser pour cela les mesures de similarités présentées dans le modèle vectoriel (voir la section 4.1).

Choix des classes

Pour la classification, le choix des classes est primordial : une classe doit être assez générale pour contenir un nombre suffisant d'objets, tout en restant la plus éloignée possible des autres classes.

Dans certains systèmes, ces classes sont définies par un expert qui s'est basé sur un échantillon de documents (Blosseville et al., 1992). Les documents sont ensuite assignés dans la classe la plus proche.

D'autres systèmes procèdent à une sélection aléatoire des initiateurs de classes, comme le système de Salton (SMART) (Salton, 1971). Ici, les documents sont pris au hasard les uns après les autres, le premier constitue une première classe, les suivants sont rattachés à la classe correspondante si elle existe (s'il existe une classe suffisamment proche du document en question), ou forment une nouvelle classe dans le cas contraire. L'attracteur d'une classe est modifié à chaque fois qu'un document lui est ajouté. Cet attracteur, aussi appelé *centroïde* est en fait la moyenne des représentations des documents contenus dans une classe.

Temps d'exécution

Le temps nécessaire pour classer un ensemble de N documents dans M classes est proportionnel à NxM. Les techniques de classification non-hiérarchiques sont donc relativement peu coûteuses en temps de calcul, et sont indiquées pour traiter des bases importantes, ou lorsque les ressources informatiques sont limitées.

4.6.2 Classifications hiérarchiques

Dans ce type de classification, on procède aux regroupements deux à deux des documents les plus proches, puis des couples de documents les plus proches, et des couples de couples de documents les plus proches et ainsi de suite jusqu'à ce qu'il ne reste plus d'éléments à apparier. Comme pour les classifications non-hiérarchiques, les mesures de similitudes sont celles développées dans le modèle vectoriel (voir la section 4.1). Le résultat de telles classifications peut être visualisé sous la forme d'un dendogramme (figure 4.2)

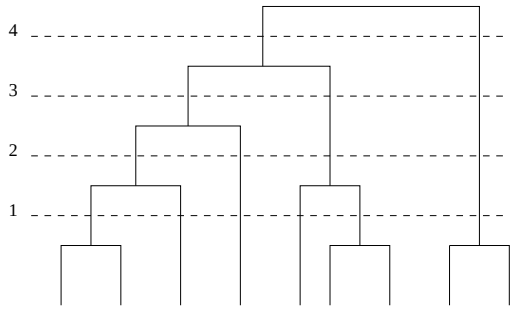

Fig. 4.2: *Dendogramme d'une classification hiérarchique*

Il existe différentes techniques de classifications hiérarchiques basées sur ces regroupements successifs (Rasmussen, 1992). Elles diffèrent surtout par la manière dont est calculée la similarité entre deux groupes de documents. Par exemple, la similarité entre deux groupes peut être associée à la similarité maximale de deux documents pris dans chacun des groupes, à la similarité minimale, ou encore à la similarité de la moyenne des documents de chaque groupe.

Temps d'exécution

Il existe un nombre important d'algorithmes pour les méthodes de classification hiérarchiques, pour lesquelles le temps nécessaire pour traiter N documents est généralement proportionnel à N^2. Ces méthodes sont donc plus lourdes en temps de calcul que les méthodes non-hiérarchiques, ce qui peut être limitatif lors de l'utilisation de très grosses bases documentaires.

4.6.3 Autres méthodes de classification

D'autres méthodes de classification ont été développées parallèlement à celles que nous venons de passer en revue.

Par exemple, la classification hiérarchique par éclatement de groupes part d'une classification non-hiérarchisée et procède à une division en sous-groupes, en fonction de différents

critères qui peuvent être le nombre maximal d'objets, ou encore l'éloignement maximal des objets dans une sous-classe.

Il existe également des systèmes de classification basés sur les réseaux inférentiels (représentations des dépendances entre les termes et entre les termes et les documents). Le système AutoClass (Ceeseman, 1988), par exemple, adopte une méthode itérative d'optimisation pour trouver le nombre optimal de classes et leurs caractéristiques pour ensuite calculer la probabilité d'appartenance des documents à chaque classe.

Les cartes auto-organisatrices (Kohonen, 1995) permettent quant à elles une classification bidimensionnelle des objets d'une base : les zones voisines des cartes contiennent des objets de caractéristiques voisines.

4.6.4 Choix d'une méthode de classification

Le but de la classification des documents est d'augmenter l'efficacité des recherches. Cette amélioration peut se retrouver sur deux plans :

- la rapidité de réponse aux requêtes : les requêtes ne sont plus comparées à chaque document, mais au centroïde de chaque classe ;
- les classes de documents correspondant aux requêtes contiennent souvent des documents qui n'auraient pas été retrouvés sans classification : les documents proches du centroïde de la classe, mais contenant peu des termes de la requête (quand la requête porte sur un thème précis, et le centroïde est plus général). Ces documents supplémentaires peuvent donner de nouveaux points de départ pour les recherches des utilisateurs.

Le choix de la classification peut se faire en fonction du point que l'on désire favoriser.

Par exemple, comme les méthodes de classification hiérarchique comportent plusieurs niveaux de classes (en fonction du niveau où on se situe sur le dendogramme, figure 4.2), le système de repérage de la classe la plus proche d'une requête est plus lourd à mettre en œuvre que dans le cas d'une classification non-hiérarchisée (qui ne comporte donc qu'un seul niveau de classification). En revanche, le niveau de classe d'une classification hiérarchisée constitue un paramètre de recherche intéressant qui permet aux utilisateurs de choisir avec quelle précision ils désirent faire leur recherche. Un utilisateur ayant une idée vague de ce qu'il recherche choisira un niveau élevé où les classes sont peu nombreuses et assez générales (le niveau 4 de la figure 4.2 par exemple, qui comprend deux classes), et vice-versa.

Un autre paramètre à prendre en compte est celui de la complexité de la mise à jour de la base. Les bases de données textuelles sont presque immanquablement amenées à recevoir de nouveaux documents (et parfois à en perdre) et un processus de mise à jour qui nécessite de recalculer une nouvelle classification à chaque évolution de la base peut être rédhibitoire à cause des ressources informatiques que cela peut demander. Relativement peu de travaux ont été faits dans ce domaine, mais on peut toutefois citer Can et Azkarahan (1990) qui proposent une stratégie pour la mise à jour des classifications évitant d'obtenir des classes distordues.

C'est quelquefois la part d'aléatoire que comportent certaines méthodes de classification, qui peut être déterminante. En effet, le résultat d'une classification non-hiérarchique peut dépendre du choix du nombre de classes, de la taille des groupements, ou encore de l'ordre dans lequel les documents sont ajoutés à la classification (dans le cas où les classes sont définies pendant le déroulement de la classification). Les méthodes hiérarchiques ne

présentent pas cet aspect et sont donc recommandées lorsqu'une classification "fiable" est recherchée (Malagnini, 1983).

Pour finir, n'oublions pas l'aspect interface homme/machine du système qui ne doit pas être trop complexe pour une utilisation aisée, tout en proposant un maximum de possibilités de recherche aux utilisateurs. Par exemple, les cartes auto-organisatrices sont bien adaptées par essence à la consultation d'une base par des utilisateurs qui n'ont pas une idée très précise des documents qu'ils recherchent. Il est possible, grâce à ce type de classification, d'affiner ses recherches en se déplaçant naturellement d'un thème à l'autre sur la carte et de trouver ce que l'on recherche sans formuler de requête (Poinçot et al., 1998).

4.7 Conclusion

Nous venons de passer en revue les principaux types de SRI, chacun étant suivi de son cortège d'avantages et d'inconvénients.

En ce qui nous concerne, nous nous sommes orientés vers les cartes auto-organisatrices (SOM) en raison des possibilités de classification et de visualisation qu'elles offrent (voir 2.2.1.2, 3.4.3.1 et 4.3.1). Ainsi, notre but est d'élaborer un système de recherche et de visualisation de données textuelles à la fois puissant et simple d'utilisation ; nous pensons que les caractéristiques des SOM peuvent nous y aider.