Société française
des sciences de l'information
et de la communication (SFSIC)

Émergences et continuité dans les recherches en information et communication

Actes du XII[e] Congrès national des sciences
de l'information et de la communication
UNESCO (Paris), du 10 au 13 janvier 2001

Traitement des données hétérogènes et formelles : vers une approche non-métrique de l'analyse du dossier de veille

Valérie Léveillé
Hervé Rostaing
Université d'Aix-Marseille III
Membres du CRRM

Introduction

Dans le contexte actuel de la société de l'information, toute organisation doit disposer d'une information pertinente et fiable pour la prise de décision. C'est notamment la mission du Système d'Information de Veille (SIV) que de fournir aux décideurs une information à forte valeur ajoutée, renseignant ces derniers sur l'évolution de leur environnement scientifique, technique, technico-économique et concurrentiel (Jakobiak, 1991 ; Dou, 1994). À partir de données collectées de différentes sources [1], l'étape d'analyse-validation doit permettre de délivrer une information élaborée, utilisable dans un processus de prise de décision. Pour cela, deux approches complémentaires sont utilisées :

- L'approche métrique qui relève d'une analyse statistique des données,
- L'approche non-métrique qui confronte directement experts et spécialistes du domaine avec les données brutes à analyser.

C'est sur cette phase critique du SIV que nous avons focalisé notre étude. Nous nous sommes plus particulièrement intéressés à l'approche non-métrique et notamment sur les moyens, les types d'organisation facilitant l'accès aux données brutes par les experts.

Après avoir exposé les apports de l'approche non-métrique dans l'analyse des données, nous présenterons comment un système basé sur une navigation de type hypertexte peut aider les experts dans leur analyse du dossier. Cette réflexion théorique s'est concrétisée par la réalisation d'un outil informatique que nous présenterons dans la dernière partie de cette communication.

1 L'ensemble des données collectées pour un sujet de veille précis constitue le « dossier de veille ».

Contexte général de l'analyse des informations dans le cadre du SIV

1.1. Approche métrique de l'analyse des données scientifiques et techniques du SIV

L'approche métrique est couramment utilisée, dans les SIV, pour analyser des volumes relativement importants de données tant scientifiques que techniques (Rostaing, 1996). Ce mode d'analyse se rapporte aux différentes techniques d'analyse statistique des données textuelles : bibliométrie, infométrie, *text-mining*. Il s'agit de déterminer, à partir de données issues de différentes sources, l'information stratégique sur laquelle se fonde la décision. Ces techniques sont basées sur l'utilisation de croisement entre informations, l'information utile n'étant pas un accroissement d'information mais une réduction d'information par des regroupements. Ceux-ci sont définis par l'ensemble des relations que les données entretiennent entre elles. Cette approche fournit une information agrégée permettant une vision synthétique de l'ensemble des documents.

Différents auteurs issus des sciences de gestion ont montré les risques qu'il y avait à baser sa décision sur une information quantitative obtenue essentiellement par agrégation de données numériques ou textuelles. Mintzberg (1994) souligne ainsi que tout processus stratégique qui s'appuie de façon excessive sur des données quantitatives peut être sérieusement biaisé et déformé. Une mise en scène des données, des résultats d'analyse, peuvent notamment orienter l'interprétation. L'information agrégée offre une vision parfois trop réductrice et imprécise des données analysées. Mintzberg insiste sur cette notion d'imprécision en montrant que le décideur perçoit la forêt alors que les opportunités se trouvent cachées sous les feuilles des arbres. Il est alors important de pouvoir accéder aux données brutes pour détecter plus facilement ces opportunités de développement. Ceci relève d'une approche non-métrique utilisée en complément de cette approche métrique de l'analyse des données.

1.2. L'approche non-métrique appliquée à l'analyse des données scientifiques et techniques du SIV

L'approche non métrique, telle qu'elle est présentée par Ndiaye et Link-Pezet (1995), est orientée vers la définition du passage de la donnée à l'information par rapport à l'utilisateur. Le passage du signe (la donnée) au signifiant (l'utilisateur) apparaît comme la résultante des processus cognitifs tels que l'accommodation, l'assimilation et la représentation. Ce processus dépend étroitement des modèles mentaux qui touchent essentiellement aux connaissances antérieures, au contexte d'information (buts, intention…) et à la représentation du problème. L'interprétation des données se fera d'autant mieux que les données seront structurées et mises en forme selon les attentes de l'utilisateur. Le processus d'assimilation des données consiste en une mise en relation de ces données, les unes avec les autres, mais également avec l'ensemble de concepts acquis par l'individu avant de commencer sa recherche. Le contenu informationnel et la pertinence d'un corpus de données éparses dépendent donc du sens qui émerge des relations mises en évidence entre les données obtenues à partir d'un traitement « computationnel » (au sens informatique) et / ou cognitif. L'information obtenue ne devient information qu'aux yeux du récepteur du message qui l'intègre dans son réseau de connaissances et de croyances (Link-Pezet, 1999).

Les systèmes issus de ce courant (comme par exemple les approches hypertextuelles) tentent de prendre en compte les éléments de stratégies interprétatives mises en œuvre lors de l'activité de l'information.

L'information utile pour la décision est lacunaire. L'intelligence stratégique, la veille stratégique supposent de reconstituer une configuration globale d'un événement ou d'un phénomène à partir d'indices fragmentaires. Cela nécessite d'étendre la signification de la donnée en la rapprochant des autres données par différents modes de traitement. Favoriser l'analyse non-métrique du dossier de veille par les experts implique, en partie, de favoriser cette mise en relation de données éparses et hétérogènes. Comment structurer et faciliter cette mise en relation ?

1.3. Organisation du dossier de veille

Boland (1987) montre qu'on ne peut résumer l'information à des données structurées. Le sens de l'information est le résultat de l'interprétation des données par le récepteur. La structuration des données ne doit pas être rigide et figée mais au contraire elle doit faciliter l'interaction entre les données et le récepteur. L'utilisateur doit avoir la possibilité « d'organiser » lui-même le corpus en fonction de la représentation qu'il a de l'ensemble des informations ainsi qu'en fonction des connaissances qu'il a du domaine étudié.

Une réflexion sur l'organisation et la présentation des informations en veille doit alors être entreprise pour favoriser cette analyse et respecter les exigences posées par cette étape, notamment au niveau de l'analyse et de l'exploration du dossier. L'expert ne doit pas seulement disposer de données agrégées, mais doit avoir la possibilité de construire son interprétation à partir de ces données en créant librement son parcours. Dans cette optique, la veille et son système d'information doivent favoriser et parfaire la chaîne de la prise de décision.

Un système d'organisation et de présentation des données doit tenir compte de différents paramètres :

- Nature multiple de l'information,
- Complémentarité de l'analyse automatique et de l'expertise des spécialistes,
- Procédés de simplification utilisés au cours de l'expertise,
- Appropriation personnelle de la donnée pour passer à l'information : chaque expert doit avoir les moyens de construire sa propre représentation du dossier de veille et construire son réseau d'informations.

La gageure des systèmes d'information est de fournir un ensemble de données structurées, mis en forme selon les besoins des utilisateurs pour faciliter et non pas remplacer l'interprétation de ceux-ci.

Organisation du dossier de veille

L'organisation des données pour l'interprétation doit alors être repensée non pas en fonction des documents eux même mais des besoins réels des utilisateurs (les experts) en matière de nature de l'information (qualitative), de modalité d'accès et de présentation de l'information.

1.4. Organisation du dossier de veille

Il existe plusieurs méthodes pour structurer et présenter les informations à l'expert. La première d'entre elles consiste à organiser le dossier de veille de façon linéaire, séquentielle. Cette organisation est de loin la plus mauvaise et reflète plus un défaut d'organisation qu'une organisation proprement dite.

Pourquoi cette organisation est-elle si préjudiciable à l'analyse ? Nous avons rapidement montré que pour « produire » de l'information réellement exploitable, chaque fragment d'information doit être rapproché de l'ensemble des informations par les liens que ce fragment entretient avec ce réseau. Or une lecture linéaire de l'ensemble de ces fragments d'information ne permet pas d'appréhender la complexité des relations du réseau.

L'organisation des informations pour l'interprétation doit être pensée non pas en fonction des documents eux-mêmes mais des besoins réels des utilisations en matière de nature de l'information (qualitative), de modalité d'accès et de présentation de l'information.

1.5. Organisation hypertexte des données

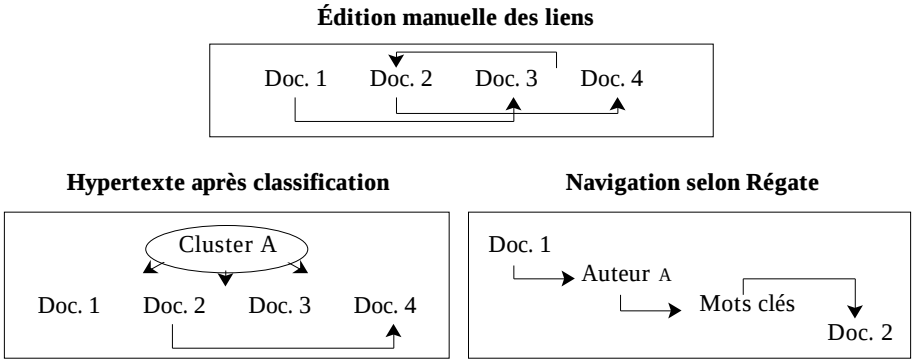
À l'inverse d'une organisation séquentielle des informations, peu apte selon nous à favoriser l'analyse de documents bruts, la structuration des données selon un modèle hypertexte favorise la découverte et offre à l'utilisateur les moyens d'accéder à toutes les informations : celles qu'il s'attend à trouver et les autres qui lui sont plus étrangères. L'intérêt des hypertextes est double :

- Faciliter la recherche d'information en permettant l'accès à un ensemble étendu de documents,
- Favoriser l'émergence de significations en fournissant l'accès aux documents originaux et en induisant une démarche intellectuelle de l'utilisateur à partir de ces éléments.

Dans ce type de structuration, il est aisé d'orienter l'analyse de l'expert en privilégiant une organisation, un parcours particulier de l'hypertexte. C'est de la façon dont est organisé l'hypertexte que dépend l'analyse. Il est donc important de construire l'hypertexte le plus objectivement possible, ce qui exclue une création manuelle de l'hypertexte. Tout d'abord, pour construire l'hypertexte et relier entre elles les informations brutes, le concepteur doit réaliser une pré-analyse des données contenues dans le dossier. C'est la vision du dossier qu'en a le concepteur (donc comment il a pré-analysé l'ensemble des données) qui sera proposée à l'analyse de l'expert. L'analyse et l'interprétation des données de la part du concepteur de l'hypertexte induiront de ce fait le travail des experts. D'autre part, le concepteur doit obtenir une vision globale de l'ensemble des documents, des liens entre les données. Le temps mis pour obtenir cette vision d'ensemble et créer ces liens serait beaucoup trop important dans un processus de veille qui doit être rapide.

La conception manuelle de l'hypertexte doit donc être abandonnée dans le cadre de notre problématique tant pour des raisons d'objectivité dans la présentation des informations que pour des raisons de délai de traitement. La figure 1 recense quelques-uns des modes de construction d'hypertexte.

Figure 1. Différents modes de conception d'hypertexte



La génération automatique d'hypertexte à partir d'une classification peut, dans certains cas limiter la navigation à une dimension : les descripteurs (qu'ils soient obtenus par extraction du champ descripteurs ou qu'ils soient issus d'une indexation automatique) ou à nombre relativement limité d'entités. La structuration hypertexte intervient ici surtout pour aider l'utilisateur dans sa recherche d'information. L'aide à la création d'informations endogènes à partir d'une structuration hypertexte est alors minimisée.

Régate, outil d'aide à l'analyse des experts

1.6. Principe de fonctionnement de Regate

Ces différentes considérations nous ont amenés à concevoir un nouvel outil d'aide à l'analyse des experts. Avec *Regate,* nous avons cherché à offrir à l'utilisateur une navigation entre les données en veille, exploitant au mieux les possibilités relationnelles de ces documents, sans limiter cette navigation à un ensemble prédéterminé d'entités. Cette vision de la conception de l'hypertexte permet de tirer profit des particularités de chaque type de bases de données. En effet, les entités permettant de mettre en relation deux brevets sont très différentes de celles qui lient deux documents scientifiques issus des bases Pascal ou Inspec par exemple. Régate propose deux niveaux de navigation :

– *Niveau de document :* les documents sont liés les uns aux autres par la présence d'un même élément. L'utilisateur a alors la possibilité de choisir la dimension selon laquelle les documents seront mis en relation : auteur, affiliation, code de classification documentaire…

– *Niveaux des formes :* nous ne nous limitons pas l'utilisation des seuls descripteurs. L'expert, dans son analyse, peut utiliser toutes entités, préalablement définies : auteurs, affiliation, périodique… Ces entités d'information sont nommées sous le terme de *rubrique*. À ce niveau, les liens sont de deux types :

    a) Document vers Formes A, Formes A étant l'ensemble des formes de la rubrique A appartenant au document : l'ensemble des auteurs associés au document, par exemple.

b) Forme 1 vers Forme A, Formes A étant l'ensemble des formes de la rubrique A présentes dans les mêmes documents que Forme 1 : par exemple, l'ensemble des descripteurs associés à un auteur particulier.

En proposant à l'utilisateur des possibilités de navigation et surtout de visualisation des associations entre les champs, Régate ne se limite pas à la stricte recherche et à la découverte d'informations, mais intègre également des fonctionnalités d'aide à la création d'informations endogènes. Accéder directement aux descripteurs associés à un auteur particulier renseigne rapidement l'expert sur les thématiques de recherche de cet auteur. Pour obtenir la même information à partir d'un hypertexte de type document vers document, l'utilisateur doit visualiser l'ensemble des documents de cet auteur et rechercher à l'intérieur de chaque document les éléments qui l'intéressent.

1.7. Navigation sous Regate

Alors que l'hypertexte, effectue une relation 1–1 entre les documents, Régate permet de visualiser des relations 1–n entre des éléments extraits de documents : les formes. Il est possible de visualiser directement les associations par exemple entre auteurs, entre auteurs et descripteurs en encore entre auteurs et pays de publications. L'interface permet d'accéder très facilement aux documents écrits par un auteur particulier ou comportant un descripteur précis.

La présentation des informations sous une forme hypertextuelle permet à l'expert, lorsqu'il parcourt le dossier, de rompre avec la linéarité habituelle des documents ainsi que la distinction formelle entre les différents types de documents : scientifique, technique, technico-économique… La navigation par association que propose Régate apporte aux experts de nouveaux modes de recherche d'information et d'apprentissage. La recherche d'information ne se fait plus uniquement à partir des éléments dont l'expert a connaissance mais par les concepts qui sont réellement présents dans la base.

L'expert doit s'impliquer dans la création d'informations élaborées, faire des choix de navigation, expliciter les liens qu'il a activé. C'est de cette explicitation que peuvent émerger des informations endogènes.

1.8. Exemple de navigation. Parcours du dossier de veille

Nous avons choisi d'illustrer notre propos par une étude réalisée dans le domaine de l'isolation thermique. 644 documents scientifiques et techniques ont été collectés sur cinq bases de données différentes Compendex, Energy Sci Tech, Jicst, NTIS et Pascal.

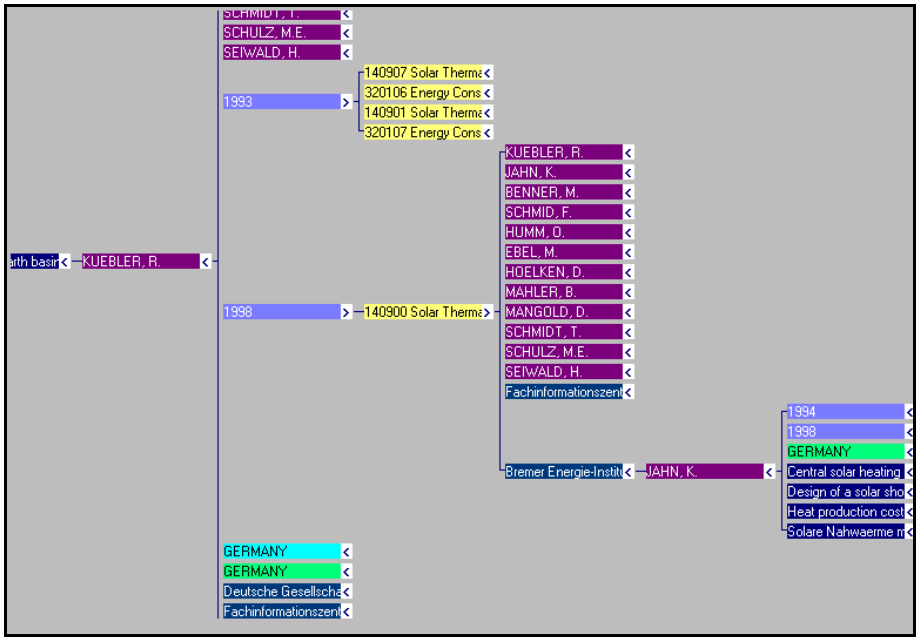
La Figure 2 présente un exemple de navigation pouvant être réalisé sous Régate. L'utilisateur a choisi de débuter la navigation à partir de l'auteur *Kuebler, R*, celui-ci étant l'un des auteurs les plus fréquents du corpus. L'utilisateur a donc choisi de détailler les éléments liés à cet auteur.

À partir de l'auteur *Kuebler, R* (1), l'expert peut connaître les co-auteurs associés à Kuebler (2), ainsi que les dates de publication, le pays de publication, les organismes associés à ce même auteur (3).

Puis, l'utilisateur a cherché à détailler les travaux réalisés en 1998 et 1993 par Kuebler (4). Dans ce cas, on affine la navigation. Puis à partir du code Energy Sci

Tech *(140900 Solar thermal insulation)* (5), l'utilisateur « ouvre » la navigation en cherchant à savoir quels sont les autres auteurs qui travaillent sur ce même domaine (représenté par ce code). Nous aboutissons à l'auteur *Jahn, K* (6), qui n'était pas associé précédemment à Kuebler, R. Nous accédons aux documents réalisés par cet auteur, ainsi qu'à d'autres données le concernant (date de publication, pays d'affiliation, organismes) (7). À tout moment de la navigation, l'utilisateur accède aux documents complets dont sont extraits ces éléments (par un simple clic de la souris).

Figure 2. Exemple de navigation sous Régate



Une présentation plus détaillée des fonctionnalités de cet outil est disponible dans la thèse *De l'organisation des données dans les systèmes d'information. Réalisation d'un outil de gestion de données hétérogènes et formelles appliqué à la veille technologique* (Léveillé, 2000).

Conclusion

Les informations collectées dans le système de veille sont incertaines, parcellaires et proviennent de sources diverses. Différentes méthodes complémentaires ont été développées pour favoriser la génération d'informations endogènes. Une partie de ces méthodes est basée sur une agrégation des données textuelles. Les résultats de ces traitements doivent être complétés par une analyse des documents eux-mêmes, menée le plus souvent par des experts du domaine. Cette interprétation du dossier ne peut provenir que d'une mise en relation des données présentes dans le corpus. Le dossier doit alors être organisé pour favoriser cette mise en relation. L'analyse du processus permettant de passer de la donnée à l'information ainsi que des méthodes à mettre en œuvre pour faciliter ce processus a permis la conception et le développement d'un nouvel outil d'aide à l'interprétation, Régate.

Cet outil offre une représentation synthétique de l'information par compilation des données présentées sous forme d'arbre. Ainsi, très rapidement, en quelques actions de souris, nous avons pu obtenir une connaissance synthétique des travaux de l'auteur Kuebler, R. L'utilisateur n'a pas eu à parcourir les dix publications de *Kuebler, R* pour regrouper ces renseignements qui sont présentés en même temps sur un seul écran. En présentant tous les liens entre les entités, l'utilisateur accède à toutes les données reliées à une forme particulière. Cette exhaustivité tend à limiter la tendance des utilisateurs à ignorer certaines informations allant à l'encontre de leur perception du dossier.

En structurant le dossier de façon objective – en fonction des liens qu'entretiennent réellement les données entre elles et non pas en fonction de la perception du dossier que pourrait en avoir un concepteur – Régate permet un accès guidé aux informations tout en laissant à l'expert les moyens de naviguer selon des éléments nouveaux qu'il découvre dans la base. Régate s'inscrit dans une perspective de mise en place d'une plate-forme de systèmes destinés à la veille assurant les fonctions de collecte de l'information, analyse statistique et enfin présentation du dossier de veille. À ce titre, il nous paraît important d'orienter nos recherches vers une meilleure prise en compte de l'information informelle.

Bibliographie

Agosti M., Melucci M., Crestani F. « TACHIR : a tool for automatic construction of hypertexts for information retrieval ». In *RIAO 94 Intelligent Multimedia Information Retrieval Systems and Management*, Rockefeller University, New-York, USA, October 11-13, 1994, p. 338-357

Boland R. « The information of information systems ». In R. Boland and Hirscheim (eds). *Critical issues information systems research.* Chichester : Wiley, 1987

Dou H. *Veille technologique et compétitivité. L'intelligence économique au service du développement industriel.* Paris : Dunod, 1995, 234 p.

Jakobiak F. *Pratique de la veille technologique.* Paris : Éditions d'organisation, 1991, 232 p.

Léveillé, V. *De l'organisation des données dans les systèmes d'information. Réalisation d'un outil de gestion de données hétérogènes et formelles appliqué à la veille technologique,* Thèse en Sciences de l'information et de la communication : Université d'Aix-Marseille III, 2000. 190 p. En ligne sur `http://193.51.109.173:81/intranet/Public/memoires.html`

Link-Pezet J. « De la représentation à la coopération : évolution des approches théoriques du traitement de l'information ». *Solaris* [on-line]. 1999, nº 5 [11/03/99] En ligne sur `http://www.info.unicaen.fr/bnum/jelec/Solaris/d05/5link-pezet.html`

Mintzberg H. *Grandeur et décadence de la planification stratégique.* Paris : Dunod, 1994, 456 p.

Ndiaye S., Link-Pezet J. « Système d'information stratégique et management : concepts et modèles ». In *Actes du Colloque Les systèmes d'information élaborée*, Île-Rousse, 30 mai–2 juin 1995, pp. 501-512

Rostaing H. *La bibliométrie et ses techniques*. Coédition Toulouse : Sciences de la société, Marseille : CRRM, 1996. 131 p.