Les temps du document et la recherche d'information

Michel Beigbeder

École Nationale Supérieure des Mines de Saint-Etienne
158, cours Fauriel
F-42023 Saint-Etienne cedex 2
mbeig@emse.fr

RÉSUMÉ. *Cet article présente un panorama des liens entre recherche d'information et aspects temporels des documents. Une première analyse amène à distinguer le temps évoqué par le discours des documents et le temps de situation de ces documents dans le temps historique. Le temps de l'univers du discours doit être pris en compte dans la phase d'indexation de la recherche documentaire. Il peut être traité par extraction d'entités nommées et plus finement par une analyse de la langue pour déterminer les relations temporelles. Le traitement des informations de catalogage si elles ne suivent pas des normes très strictes est en fait un problème voisin. Le temps de publication, qui est dans le monde de l'édition traditionnelle la principale donnée de catalogage à caractère temporel, devient dans le monde du document numérique une donnée fondamentale permettant de modéliser l'évolution des documents. Nous introduisons les notions de collections « muable » et immuable. Nous évoquons aussi les questions de granularité de représentation du temps.*

ABSTRACT. *This article presents a survey of the interactions between information retrieval and temporal aspects of documents. First the time evoked by the discourse and the historic situation time of the documents themselves are distinguished. The universe of discourse time should be taken into account in the indexation step of information retrieval. It can be processed by named entities extraction and more finely by natural language processing methods in order to determine temporal relations. The processing of metadata information if the latter does not follow very strict rules in fact is a neighbour problem. The publication time, which is in the traditionnal publishing world the main temporal metadata, becomes in the digital document world a fundamental piece of information. We introduce the notion of mutable and immutable collections. We quote the questions of time representation granularity.*

MOTS-CLÉS : *recherche d'information, recherche documentaire, temps des documents.*

KEYWORDS: *information retrieval, times of documents.*

DN – 8/2004. Temps et documents, pages 55 à 64

1. Introduction

Le domaine de la recherche d'information en tant que thème de recherches informatiques est quasiment aussi ancien que l'informatique elle-même. A l'origine, il s'inscrivait dans la poursuite du travail des documentalistes dont le rôle a toujours été de classer et classifier les documents dans le but de pouvoir les retrouver à la demande d'un utilisateur. Ce modèle de recherche d'information, que nous appelerons plus précisément « recherche documentaire », reste toujours en usage aujourd'hui et est même devenu bien plus connu du grand public depuis l'avènement de la Toile sur internet, rapidement suivi par l'apparition de « moteurs de recherche » (*Wanderer* en 1993, *Lycos* et *Excite* en 1995, *Altavista* en 1996...) et de « répertoires thématiques » (*Yahoo* en 1994...). Cependant, l'aspect recherche documentaire de la recherche d'information n'est pas le seul, et traditionnellement on y retrouve aussi le « routage », le « filtrage » et l'« extraction d'information ».

Avant de présenter où des questions concernant le temps interviennent dans le domaine de la recherche d'information, nous avons à évoquer ce que le mot-clé « temps » (*time* en anglais) représente majoritairement dans ce domaine. Le modèle classique de la recherche documentaire se découpe temporellement en deux phases : la phase d'indexation et la phase d'interrogation. Le plus souvent, lorsque le mot temps est utilisé, il fait référence à la durée d'exécution par un ordinateur de la phase d'indexation ou de la phase d'interrogation. Ces temps sont fortement liés à la complexité algorithmique des méthodes mises en œuvre et à la qualité de l'ingénierie logicielle. Le temps d'interrogation, visible et palpable par les utilisateurs d'un système de recherche d'information, se doit d'être compatible avec les exigences des utilisateurs et leur nombre. De l'autre côté, la durée du temps d'indexation doit permettre au système de recherche d'information d'être en phase avec l'évolution de la collection qu'il a à indexer. Nombre de travaux en recherche d'information abordent ces aspects temporels d'un point de vue informatique, mais ce ne sont pas ceux qui nous intéressent ici.

Nous allons présenter brièvement comment différents aspects temporels interviennent dans les activités de recherche d'information. Nous ferons référence aux trois temps :

– temps de l'univers du discours ;

– temps de publication ;

– temps de l'évolution.

2. Le temps de l'univers du discours

Historiquement, le temps de l'univers du discours est le premier à avoir été explicitement pris en compte dans les domaines de la recherche documentaire et de l'extraction d'information. La prise en compte de ce temps consiste à considérer que les documents et les besoins d'information contiennent des concepts qui sont spécifiquement liés à un aspect temporel : une date, un lien avec un événement historique. Il

s'agit donc de repérer ces concepts lors de l'indexation dans le but de pouvoir spécifiquement les retrouver. Cette tâche consiste à reconnaître une certaine classe d'« entités nommées ». Les conférences MUC [1] se sont attachées à définir des tâches particulières et ont retenu trois classes d'entités nommées : celles comprenant i) les noms propres et les acronymes (*organization*, *person*, et *location*), ii) les expressions temporelles absolues (*date* et *time*) iii) les expressions numériques monétaires et de pourcentage (*money* et *percent*).

Cette reconnaissance est compliquée par le fait que les dates peuvent prendre des formes très diverses dépendant des conventions nationales. Citons par exemple l'ordre entre le numéro du mois et le quantième qui diffère dans les usages anglais et américain. Ainsi, selon l'ordre retenu, l'écriture « *3-9-2004* » désigne le 3 septembre (G.B.) ou le 9 mars (U.S.A.) de l'année 2004. Certaines applications informatiques qui doivent s'appuyer sur des dates absolues ne peuvent admettre différentes interprétations et normalisent de ce fait l'écriture qui doit être partagée par tous les agents. En particulier les agents de messagerie électroniques doivent respecter à la lettre les recommandations du RFC 2822 (2001) pour spécifier une date et une heure ; en voici un exemple « `Fri,␣␣3␣Sep␣2004␣17:01:56␣+0200␣(MEST)` ».

Dans certains cas, la reconnaissance devrait s'appuyer pour être complète sur des connaissances bien plus larges que celles des différentes conventions d'écriture des dates. Par exemple, depuis le 11 septembre 2001, l'expression « *11 septembre* » peut désigner soit la date du 11 septembre 2001 même si l'année n'est pas mentionnée, soit l'événement associé à cette date (a-t-on encore affaire à une entité nommée de type date ?) soit encore une date ordinaire.

D'autres travaux se sont intéressés à l'analyse temporelle de la langue naturelle. Les premiers remontent à plus de 20 ans (Hirschman, 1981) et se poursuivent encore aujourd'hui (Mani and Wilson, 2000 ; Mani et al., 2004). Dans ces travaux, le but est de découvrir les relations temporelles entre des événements. Ces relations sont déduites à partir d'indices comme les conjonctions (*quand*, *pendant que*, *avant que*...), le temps de conjugaison des verbes, les expressions adverbiales (*après l'accident*, la *semaine dernière*, *plus tard*, *hier*...) en plus des références absolues évoquées précédemment. Ces aspects sont plus spécifiquement pris en compte aujourd'hui par les communautés qui s'intéressent aux traitement automatique de la langue (TAL en français, NLP en anglais, *Natural Language Processing*).

L'application des méthodes d'extraction de dates et de compréhension automatique des relations temporelles trouve son application dans les systèmes Question-Réponse (Q/R en français, Q/A en anglais, *Question Answering*). Le domaine Q/R est à la frontière de la recherche d'information ; en effet la problématique va au-delà du dépistage des documents pertinents à un besoin d'information puisqu'il s'agit de répondre explicitement au besoin d'information qui est une question. Lorsque la question porte sur un aspect temporel (*Quand... ?*, *Combien de temps a duré... ?*), ces applications

1. http://www.itl.nist.gov/iaui/894.02/related_projects/muc/index.html

doivent s'appuyer entre autres sur des raisonnements, donc sur l'intelligence artificielle, et en particulier sur la partie qui traite des raisonnements temporels.

Ce temps de l'univers du discours est aussi celui qui apparaît d'une façon explicite dans des bases de données ayant un contenu où des dates (sous une forme exacte ou approchée, absolue ou relative, d'événement ou d'intervalle avec une durée, etc.) sont représentées. De nombreux travaux existent dans ce domaine. Des groupes de travail et des communautés organisent des conférences sur ce sujet (Rolland et al., 1988 ; Clifford and Tuzhilin, 1995 ; Etzion et al., 1998). Il faut noter que les aspects temporels sont souvent liés à des aspects spatiaux (Hadzilacos et al., 2003), et que ce domaine à son tour se trouve être lié à celui des systèmes d'informations géographiques (SIG).

Pour répondre au mieux à des besoins d'information qui manifestent des aspects temporels, plusieurs approches nécessitent d'être assemblées. Il faut d'abord reconnaître dans les documents les références temporelles (extraction d'entités nommées) et les relations temporelles (grâce aux méthodes de traitement automatique de la langue). Les éléments d'information recueillis permettent de renseigner des bases de données dans lesquelles peuvent intervenir des raisonnements temporels.

3. Le temps de publication

Le temps documentaire place un document parmi un réseau d'autres documents. Dans le modèle classique de l'édition, un document se positionne par rapport à des documents déjà parus, donc plus anciens que lui. C'est ici, au niveau de la date de publication (souvent seulement une indication d'année) qu'intervient une notion de temps. Si l'on revient à la recherche d'information, cela se traduit dans un besoin d'information par des critères concernant cette date de parution. Ces critères peuvent se traduire par des contraintes booléennes sur un attribut particulier – celui qui conserve la date de parution – ou un critère de classement, typiquement trier les documents retrouvés par un système booléen du plus récent au plus ancien. Le temps n'est plus alors, comme dans le cas du temps de l'univers du discours, relatif au contenu du document, mais il concerne un attribut externe au document ; on parle de *méta-information*.

Les méta-informations (*metadata* en anglais) sont un moyen d'expression des données du *catalogage* dans les bibliothèques. De ce fait, leur mise en place dans les ordinateurs a conduit à de nombreuses normalisations pour donner des possibilités d'échange et d'interrogation sur plusieurs bases de données bibliographiques. Il s'agit donc ici plus d'une problématique de normalisation que d'une problématique de modélisation du temps permettant des raisonnements ou d'une problématique d'extraction d'information. En effet, les dates en question sont non ambiguës et renseignées par des documentalistes, par contre ces derniers doivent respecter des conventions permettant le partage de l'information créée d'une façon fiable.

Toutefois la complexité de certains système de codage de méta-informations fait que la notice d'un document peut être considérée comme un document à part entière.

Le champ 250 de MARC [2] qui permet de préciser l'édition peut par exemple être rédigé ainsi :

```
250    ##$a2nd ed.
250    ##$aRev. as of Jan. 1, 1958.
```

et il peut y avoir ou non indication d'une date. Le champ 260 qui précise les informations relatives à la version de publication d'un travail peut être complexe :

```
260    ##$aBelfast [i.e. Dublin] :$b[s.n.],$c1946 [reprinted 1965]
260    ##$aWashington, D.C. (1649 K St., N.W., Washington 20006) :
       $bWider Opportunities for Women,$c1979 printing, c1975.
```

La partie contenant une indication temporelle est facilement repérable grâce à son introduction par `$c`. Par contre, son analyse complète n'est pas triviale, puisqu'elle peut contenir plusieurs dates, et différentes indications (*printing*, *reprinted* dans ces exemples).

Au niveau de la granularité de la représentation du temps, les besoins ne sont pas les mêmes dans toutes les applications. Pour un catalogage de publications traditionnelles (livres), l'année ou encore l'année et le mois suffisent. Pour le catalogage de clichés photographiques, l'heure doit aussi être précisée, et les normalisations doivent permettre de préciser l'heure utilisée : heure locale, heure GMT, ou autre. Il y a donc une question de granularité dans la représentation des instants. Mais outre les problèmes de formatage de l'écriture de la date, il faut aussi prendre en compte l'évolution de la modélisation sous-jacente du temps. Pour fixer les idées, deux exemples. Le premier se place au niveau de granularité de l'année. On considère généralement que le passage au calendrier grégorien est survenu en 1752. A cette époque, la plupart des pays avaient reconnu ce calendrier (mais quelques-uns ne le reconnurent qu'au début du 20$^{\text{ème}}$ siècle). Les onze jours suivants le 2 septembre furent éliminés, et donc le calendrier de ce mois est un peu particulier : le lendemain du 2 septembre fut le 14 septembre. Second exemple, au niveau de granularité de la seconde, il faut savoir que le *International Earth Rotation and Reference Systems Service* [3] est responsable de l'annonce de l'ajout ou du retrait d'une seconde additionnelle pour ajuster le temps UTC mesuré avec un ensemble d'horloges atomiques, et le temps astronomique, plus irrégulier. Ces secondes additionnelles sont ajoutées ou retirées à la fin du mois précédent un premier juillet ou un premier janvier. Depuis la création du système des secondes additionnelles il y a eu 22 secondes ajoutées et la dernière remonte au 31 décembre 1998. Selon les applications on peut avoir besoin ou non de tenir compte de ces irrégularités.

Autour des méta-informations, les actions sont très nombreuses et éventuellement ciblés sur des domaines d'application particulier. Le site de *International Federa-*

2. http://www.loc.gov/marc/. Les exemples qui suivent ont été extraits de ce site.
3. http://www.iers.org/iers/

tion of Library Associations and Institutions [4] régulièrement mis à jour fournit de nombreux points d'entrée dans ce domaine. On peut noter que lorsque des méta-informations temporelles existent dans les modèles de méta-informations, celles-ci sont souvent couplés à des méta-informations spatiales. C'est particulièrement évident pour les méta-informations des SIG, mais aussi dans la proposition de Dublin Core [5] où c'est un seul et même élément *Coverage* qui recouvre les aspects spatiaux et temporels.

Du côté des outils de recherche d'information, ceux des bibliothèques permettent de donner des critères booléens sur la date de publication. Une intégration plus poussée nécessiterait d'avoir une prise de décision multicritère combinant des aspects de pertinence de contenu (aspects bien connus, bien que restant difficiles en recherche d'information) avec des aspects d'adéquation plus ou moins correcte à des critères sur les méta-informations, et de ce fait s'écarter d'un modèle purement booléen pour ces critères, en particulier celui du temps. A notre connaissance ces problèmes sont peu abordés, des premières propositions étant celles de Furuta and Na (2002) dans le cadre de la navigation sur la Toile.

4. Le temps de l'évolution

Parler de l'évolution d'un document suppose de préciser ce qu'est un document et de le distinguer de son contenu. En effet, pour qu'il y ait évolution quelque chose doit changer, et si l'on considère qu'un document *est* son contenu ; deux contenus différents (ne serait-ce que par une virgule) sont différents. Une formalisation de l'évolution consiste donc à considérer que les documents peuvent être identifiés indépendamment de leur contenu.

Le temps de l'évolution des documents apparaît avec les documents numériques. En effet, sous une forme classique les documents n'ont pas d'évolution. Dans le monde des livres, il y a la notion d'édition et la notion de tirage, mais leurs instances restent en petit nombre et sont espacées dans le temps, de plus elles sont parfaitement identifiées et une nouvelle édition ne vient pas en remplacer une plus ancienne mais vient s'ajouter à la collection. Dire qu'il n'y a pas d'évolution d'un document signifie qu'une référence bibliographique ou une description dans un catalogue de bibliothèque désignent un document, bien sûr par son titre et ses auteurs, mais aussi par son année d'édition et son numéro d'édition (par exemple : 2$^{\text{ème}}$ édition, 1969) ; le tout permettant de préciser de façon unique un contenu. Tout est fait donc pour qu'il n'y ait pas de distinction entre l'identification d'un document et son contenu.

Dans le monde des documents numériques en général et de la Toile en particulier, l'identifiant d'un document est différent de son contenu. Sur un poste de travail personnel ou partagé, l'identifiant d'un document est le nom du fichier qui le supporte, et

4. http://www.ifla.org/II/metadata.htm
5. http://dublincore.org/documents/dces/

son contenu est susceptible de changer sans que le nom ne change. De même, sur la Toile l'identifiant est une URL.

Pour formaliser, nous appelons $\mathcal{I}$ l'ensemble des identifiants de la collection ; dans le cas de la Toile, $\mathcal{I}$ est l'ensemble des URL valides. Nous modélisons le temps par un axe de nombres réels $\mathbb{R}$. Enfin, nous appelons $\mathcal{D}$ l'ensemble des contenus de documents. L'évolution de la collection se modélise alors par une fonction $\delta : \mathcal{I} \times \mathbb{R} \rightarrow \mathcal{D}$. Cette fonction associe à un couple composé d'un identificateur de document x et d'une date t le contenu de ce document à cette date : $\delta(x, t)$.

Cette formalisation nous permet d'expliciter deux modèles de définition d'une collection : le *modèle muable* et le *modèle immuable*.

4.1. *Collection muable*

Une *collection muable* ne connaît qu'une instance temporelle d'un identifiant de document. Autrement dit la collection se modélise par une fonction $\tilde{\delta} : \mathcal{I} \rightarrow \mathcal{D}$. Le problème à traiter, dû à l'évolution, consiste à ce que la fonction $\tilde{\delta}$ soit aussi proche que possible de la fonction $\delta_0 : x \mapsto \delta(x, t_0)$ où t_0 désigne l'instant présent. Dans le cas d'un système de recherche d'information, pour que la fonction $\tilde{\delta}$ soit exactement δ_0, il faudrait que l'agent responsable de la modification du contenu d'un document avertisse l'agent responsable de l'indexation.

Les moteurs de recherche sur la Toile travaillent sur des collections muables, mais ils ne sont pas prévenus des modifications de contenu. Ils reconstruisent donc en permanence leur base d'index (et donc la base de documents). Cette reconstruction se fait par un parcours régulier de ce qui est connu i) pour voir si derrière la référence (l'URL) se trouve toujours un document, ii) pour charger le contenu textuel courant du document, iii) pour référencer les documents pointés avec des URL par ce document. Les moteurs essayent donc d'avoir dans leur base des index les plus à jour possibles. Les solutions sont orientées essentiellement dans la recherche de la plus grande vitesse possible de parcours et sont abordées par de l'ingéniérie logicielle sur des parcs d'ordinateurs. La mise à jour par rapport à la volatilité de la collection pose des problèmes plus difficiles si le système de recherche d'information n'est pas centralisé et qu'on se trouve donc dans un cadre de recherche d'information distribuée (Sato et al., 2003).

4.2. *Collection immuable*

Une *collection immuable* considère plusieurs versions temporelles associées à un même identificateur de document. Par contre, là encore, si l'agent responsable de la mise à jour de la collection n'est pas prévenu des modifications apportées aux documents, il n'a pas nécessairement toutes les versions. D'un point de vue pratique, maintenir la collection consiste à échantillonner dans le temps la fonction δ : on ne connaît donc pas exactement les $\delta(x, t)$ pour toutes les valeurs de t, mais seulement

pour un sous-ensemble fini (t_i). La suite (t_i) est une suite croissante de dates, bornée par la date actuelle.

Une collection immuable modélise la fonctionnalité d'archivage. Sur la Toile, l'archivage est partiellement traité par le site de INTERNET ARCHIVE [6]. Une fois une archive constituée, les questions de recherche d'information qui se posent sont celles évoquées dans la section 3.

5. L'insertion des documents dans leur réseau

Une particularité des documents numériques est que leur insertion dans le contexte des autres documents y est accessible, prenant la forme des liens entre les pages HTML, liens représentés par les balises `<A HREF="...">`. La nouveauté n'est pas tant l'existence de ces liens – ils existent dans la tradition scientifique sous la forme de références bibliographiques que l'on trouve à la fin des articles – que dans leur disponibilité au même titre que le contenu textuel.

Il y a cependant une différence, due au temps de l'évolution des contenus, entre les références au passé des citations bibliographiques traditionnelles et les liens hypertextuels entre les documents. Les premiers ne font référence qu'à du passé identifiable et immuable. Les seconds ne sont représentés que par un pointeur vers des données qui elles ne sont pas immuables.

Des travaux de plus en plus nombreux – et certains moteurs de recherche – utilisent ces liens sans toutefois aborder ce problème de volatilité. La différence que nous évoquons modifie la sémantique des méthodes qui sont utilisées. Par exemple, la méthode des cocitations introduite par Small (1973) permet de structurer un corpus par les citations qu'il fait dans son passé. L'utilisation de cette méthode sur un graphe de citations qui n'est plus seulement orienté vers le passé, comme celui de la Toile, permet de faire émerger de nouvelles interprétations (Prime-Claverie et al., 2004) où la structuration n'est plus seulement dans l'espace documentaire mais dans l'espace des acteurs qui ont participé à la publication des documents.

6. Filtrage et routage

La particularité du filtrage et du routage par rapport à la recherche documentaire est que la collection de documents est dynamique dans le sens où l'on a à gérer un flot de documents. Par contre, les besoins d'informations sont relativement statiques. Cette particularité liée au flot et donc au temps est bien évidemment présente dans tous les travaux. Nous ne donnerons donc pas de références particulières. On peut cependant insister sur un aspect particulier de ce sous-domaine qui concerne la découverte de

6. http://www.archive.org/

« nouveautés ». Une piste de TREC [7] est consacrée à cette activité depuis 2002 et permet à travers les actes de la conférence de connaître les acteurs de ce sous-domaine.

7. Conclusion

Ce rapide panorama des interactions entre recherche d'information et aspects temporels montre que les spécificités temporelles sont traitées au cas par cas. Malgré tout il y a une continuité de problématique. Le temps de l'univers du discours énonce des propositions dont il assure qu'elles sont vraies à certains instants ou pendant une certaine période. Les documents qui contiennent ces propositions sont eux-mêmes sujets à évolution au cours du temps historique, à condition de distinguer comme nous l'avons modélisé l'identifiant d'un document de son contenu. La date de publication dans le temps historique permet à un lecteur de cerner la signification des dates mentionnées dans l'univers du discours, et au-delà lui permet de replacer tout le contenu textuel dans le cadre culturel de son époque d'écriture.

La recherche documentaire connaît bien les problèmes de polysémie, et cette dernière est toujours augmentée par l'accroissement du volume des collections. Les concepts évoluent au gré des évolutions des civilisations, le vocabulaire évolue lui aussi, mais certains mots sont réutilisés avec des sens différents, toute l'étymologie est là pour nous le rappeler. Toutefois la relative jeunesse des documents électroniques fait que la très grande majorité des documents disponibles sous forme numérique sont contemporains, si bien qu'il n'y a pas eu d'études prenant en compte en recherche documentaire l'évolution du sens des mots.

8. Bibliographie

Clifford J., Tuzhilin A., editors. *Recent Advances in Temporal Databases, Proceedings of the International Workshop on Temporal Databases, Zürich, Switzerland, 17-18 September 1995*, Workshops in Computing, 1995. Springer. ISBN 3-540-19945-4.

Etzion O., Jajodia S., Sripada S. M., editors. *Temporal Databases : Research and Practice, (Dagstuhl Seminar, June 23-27, 1997)*, volume 1399 of *Lecture Notes in Computer Science*, 1998. Springer. ISBN 3-540-64519-5.

Furuta R., Na J.-C. Applying cat's programmable browsing semantics to specify world-wide web documents that reflect place, time, reader, and community. In *ACM Symposium on Document Engineering*. ACM, 2002.

Hadzilacos T., Manolopoulos Y., Roddick J. F., Theodoridis Y., editors. *Advances in Spatial and Temporal Databases, 8th International Symposium, SSTD 2003, Santorini Island, Greece, July 24 - 27, 2003*, volume 2750 of *Lecture Notes in Computer Science*, 2003. ISBN 3-540-40535-6.

7. http://trec.nist.gov/

Hirschman L. Retrieving time information from natural-language texts. In Oddy R. N., Robertson S. E., van Rijsbergen C. J., Williams P. W., editors, *Information Retrieval Research, Proc. Joint ACM/BCS Symposium in Information Storage and Retrieval*. Butterworths, 1981. ISBN 0408107758.

Mani I., Pustejovsky J., , Gaizauskas R., editors. *The Language of Time : A Reader*. Oxford University Press. A paraître.

Mani I., Wilson G. Temporal granularity and temporal tagging of text. In *AAAI-2000 Workshop on Spatial and Temporal Granularity, Austin*, 2000.

Prime-Claverie C., Beigbeder M., Lafouge T. Transposition of the co-citation method with a view to classifying web pages. *Journal of the American Society for Information Science and Technology (Special Issue on Webometrics)*, 55(14) :1282-1289, 2004.

RFC 2822. Internet message format, 2001. URL http ://www.faqs.org/rfcs/rfc2822.html

Rolland C., Bodart F., Léonard M., editors. *Temporal Aspects in Information Systems, Proceedings of the IFIP TC 8/WG 8.1 Working Conference on Temporal Aspects in Information Systems, Sophia-Antipolis, France, 13-15 May, 1987*, 1988. North-Holland / Elsevier. ISBN 0-444-70373-X.

Sato N., Uehara M., Sakai Y. Temporal information retrieval in cooperative search engine. In *DEXA Workshops*, pages 215-220. IEEE Computer Society, 2003.

Small H. Co-citation in the scientific literature : a new measure of the relationship between two documents. *Journal of the American Society for information Science*, 24(4) :265-269, 1973.