



HAL
open science

Bibliométrie et linguistique : Évaluation de la production scientifique et annotation sémantique

Marc Bertin, Jean-Pierre Desclés, Brahim Djioua, Yordan Krushkov

► To cite this version:

Marc Bertin, Jean-Pierre Desclés, Brahim Djioua, Yordan Krushkov. Bibliométrie et linguistique : Évaluation de la production scientifique et annotation sémantique. 9e Colloque International sur le Document Electronique (CIDE.9) dans le cadre de la Semaine du Document Numérique (SDN'06), Sep 2006, Fribourg, Suisse. hal-01954184

HAL Id: hal-01954184

<https://hal.science/hal-01954184>

Submitted on 13 Dec 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Bibliométrie et linguistique : Évaluation de la production scientifique et annotation sémantique

Marc Bertin¹, Jean-Pierre Descles², Brahim Djioua³, Yordan Krushkov

*Laboratoire LaLICC
(Langage, Logique, Informatique, Cognition et Communication),
28 rue Serpente, 75006 Paris, France
Université Paris-Sorbonne/CNRS UMR8139*

¹marc.bertin@paris4.sorbonne.fr

²jpdescles@paris4.sorbonne.fr

³bdjioua@paris4.sorbonne.fr

Résumé:

L'identification et l'évaluation de la production scientifique sont un problème d'actualité. Nous nous sommes donc intéressés à étudier comment les auteurs sont cités à travers les publications scientifiques. Cette approche linguistique permet de catégoriser les renvois bibliographiques d'une publication. L'application informatique de cette étude s'appuiera sur la plateforme informatique EXCOM (EXploration COntextuel Multilingue).

Abstarct:

The identification of the scientific production and the evaluation of the researchers is a problem of current events. It is the reason for which we suggest being interested in the way are quoted the authors in articles. This approach allows to categorize quotation and annotate automatically a text. The computer application of this study will be integrated into the platform EXCOM (Multilingue Contextual EXploration).

Mots-clés: Evaluation, EXCOM, exploration contextuelle, linguistique, bibliographie

Keywords: Evaluation, EXCOM, contextual exploration, linguistic, bibliography

Introduction

Identifier la production scientifique, tout comme l'évaluation de la science en générale, sont des exercices périlleux. Sans rentrer dans un débat, je soulignerai une prise de conscience plus vive de cette problématique avec la classification de Shangaï. Aussi, si l'approche de Garfield est loin d'offrir à travers le Facteur d'Impact une solution très pertinente, elle n'en reste pas moins la solution la plus présente, et cela malgré les biais introduits et souvent discutés dans la littérature. Nous allons donc essayer de comprendre le pourquoi de cette situation et proposer

une approche nouvelle sans nous appuyer sur l'approche statistique habituelle, c'est-à-dire celle de Zipf qui utilise les fréquences d'apparitions de termes dans le texte, mais en préconisant l'approche linguistique. En effet, les méthodes statistiques usuelles comme Okapi et autres "tf.idf" sont largement utilisées dans les systèmes de recherche d'informations et d'évaluation. Nous proposerons donc une classification des relations entre auteurs à base de critères qualitatifs basée sur une étude linguistique textuelle.

Origine de l'approche statistique

Nous prendrons comme point de départ la bibliologie telle qu'elle est définie par Otlet. Elle décrit la science systématique et raisonnée du livre qui a pour objet l'histoire du livre et ses procédés de fabrication, de diffusion et de conservation. L'application de l'outil statistique à l'univers de la bibliologie conduit naturellement à la bibliosociométrie qui est la mesure de l'activité du livre et du document sur l'homme et la société, et à la bibliométrie, qui est l'ensemble des méthodes et techniques quantitatives, de type mathématique et statistique, susceptible d'aider à la gestion des bibliothèques et d'une manière très générale des divers organismes ayant à traiter de l'information. Il s'agit là de la définition proposée par A. Pritchard. En fait, il est clair que le livre dépend de par sa matière de l'économie et de par son texte, il relève de la linguistique et de la textologie.

Société de l'information

À la source des sciences de l'information, nous pouvons citer trois lois qui sont celles de Bradford, Lotka et Zipf. Nous connaissons aussi cette dernière sous le nom de Zif-Pareto, soulignant ainsi le lien avec le monde de l'économie. La loi de Lotka est relative à la production d'articles par les chercheurs. Celle de Zipf exprime la fréquence d'apparition des mots dans un texte. Quant à Bradford, elle exprime la répartition des articles dans les revues. C'est à partir de la loi de Lotka que Price proposera l'idée que le nombre d'auteurs les plus productifs est donné par la racine carrée du nombre total d'auteurs. Malgré les travaux de Glänzel et Schubert, il fut difficile de proposer une formulation mathématique en concordance avec les données empiriques. Cependant, la grande idée de Price a été de définir une propriété essentielle du champ scientifique en proposant la Distribution sur les Avantages Cumulés. Cette théorie se propose comme générale et unificatrice des lois empiriques de la bibliométrie. L'idée sous-jacente est qu'une minorité de scientifiques se trouve être à l'origine de la majorité des publications dans un domaine. Nous citerons Price [PRI63] afin de mettre en évidence sa formation de physicien et donc son approche :

« On étudie le comportement d'un gaz à différentes conditions de température et de pression. On ne s'intéresse pas à une molécule appelée Georges, se déplaçant à une vitesse spécifique et située en un endroit spécifique à un instant donné; on considère seulement la moyenne de l'ensemble total des molécules où certaines sont plus rapides que d'autres, où elles sont situées au hasard et se déplaçant en différentes directions. »

Théorie de la citation

Au delà de l'aspect purement quantitative de cette thermodynamique des auteurs, le questionnement sur les motivations des auteurs à citer et la signification de la citation a permis de mettre en évidence deux écoles de pensée. L'étude des citations nécessite de comprendre les normes utilisées, les différentes fonctions de la citation, de leurs qualités ainsi que les motivations et les raisons pour citer des travaux. On peut citer [CRO84], [KIN87], [LIU93] et [LEY98]. La fonction communicative de la citation se résume à deux courants. La première approche peut se définir par une citation de Wilson (1999):

« Document is cited in another document because it provides information relevant to the performance and presentation of the research, such as positioning the research problem in a broader context, describing the methods used, or providing supporting data and arguments. »

L'auteur qui cite est conditionné par les normes de la science en générale, et plus particulièrement par les normes de son domaine de recherche. Cela rejoint les points de vue de Garfield, Price et Cole [PRI63], [PRI65] et [COL92]. Il est admis dans cette théorie que les citations sont égales entre elles et qu'elles sont suffisantes à l'argumentation de l'auteur. Une des idées fortes de Price est qu'il a été le premier à souligner la possibilité de mettre en relation les auteurs afin de pouvoir cartographier la science ou au moins un domaine de celle-ci à travers une analyse des co-citations.

S'opposant à l'approche à la théorie normative, l'approche sociale constructiviste prône que les citations sont des instruments rhétoriques afin de persuader les lecteurs selon des critères autres que scientifiques. On peut renvoyer pour cela aux écrits de [LAT87]. Un travail de synthèse, en marge de cette problématique, détail plus en profondeur cette réflexion [SCH04] dans son introduction.

Les autres travaux

Cependant, différentes études ont été menées à des fins plus applicatives comme le résumé automatique. Il n'est plus à démontrer l'importance des citations qui sont très présentes dans les systèmes informatiques comme le SCI de l'ISI ou CiteSeer. Garfield a proposé le premier en 1955 un système d'indexation des citations. Les hyperliens offrent la possibilité de naviguer d'un article à l'autre avec facilité et permettent ainsi d'identifier plus facilement les travaux connexes au domaine de l'article. L'étude des citations n'est donc pas nouvelle et différentes catégorisations ont vu le jour. Celles-ci remontent aux travaux de Garfield [GAR65] ou de Lipetz [LIP65]. Nous pourrions donner comme exemple les catégories suivantes : *conceptual or operational, organic or perfunctory, evolutionary or juxtapositional, and confirmatory vs. negational* tel que proposées par [MOR75, MUR75]. White proposera une classification basée sur des études sociologiques et déterminera les raisons pour lesquelles certains travaux sont mis en avant ou plus cités que d'autres [WHI04].

Bradshaw construira une nouvelle métrique RDI (pour Reference Directed Indexing) [BRA03]. De façon succincte, la pondération variera en fonction des termes présents dans la « citation » qui est un néologisme anglophone indiquant le segment textuel entourant la citation. Cette méthode recherche dans l'article cité les termes trouvés dans le segment textuel. Cela permet de pondérer plus fortement les articles cités ayant les mêmes termes.

Nanba identifie les citations à des fins de résumé. Son hypothèse de travail est qu'une citation représente un résumé succinct selon le point de vue de l'auteur. [NAN00 et al.] repère des segments textuels (*citations area*) pour identifier les citations afin de générer des résumés. Nous soulignerons que dans ces travaux, il propose également une catégorisation des citations (*citations types*) afin d'organiser les « aires de citations ».

Plus proche de nos préoccupations, nous pouvons citer les travaux de Simone Teufel qui propose également une classification [TEU00, MOE00], [TEU01]. Sa catégorisation, en 13 points, couvre assez largement les citations et leurs motivations. De même, Bonzi s'est intéressée à la citation négative expliquant qu'une citation n'était pas forcément un signe d'acceptation de la part de celui qui cite [BON82]. Enfin, d'un point de vue de l'automatisation, les travaux de Mercer [MER04, MAR04] proposent d'utiliser ces classifications qui sont ignorées par les systèmes d'indexation de citation.

L'ensemble de ces travaux sur les catégorisations n'offrent pas de solutions de cartographies à grande échelle. De plus elles sont limitées dans le cadre d'une application ou d'une méthode. Elles ne s'appuient pas sur une étude linguistique qui permettrait de traverser les domaines en s'intéressant non pas aux termes clés mais aux relations entre les termes. C'est l'approche linguistique que nous prônons ici.

Limitation de l'approche quantitative

Mesurer la qualité de la production est relativement difficile dans le sens où les indicateurs bibliométriques caractérisent le contenant et non le contenu. Elle apporte une valeur et des mesures, mais ils ne sont pas et ne doivent pas être des signes de la qualité de la recherche scientifique. On constatera ces dernières années une attitude du « publish or perish » conduisant à des pratiques d'écriture qui peuvent mettre en péril la qualité des articles. Cela peut provoquer des comportements antiscientifiques comme le plagiat, la publication dans une revue où le FI est élevé plutôt que dans une revue adéquate ou bien encore de diviser les données en parties ridiculement petites. L'un des risques encouru par cet état des faits est sans doute à court terme une production scientifique accrue, mais d'une qualité moindre, nécessitant de parcourir un certain nombre de publications pour couvrir une pensée ou un concept. À moyen terme un risque d'uniformisation de la recherche est présent. Cette homogénéité a déjà été soulignée et plusieurs articles présentent les biais introduits par cette méthode d'évaluation. Au-delà de l'aspect rédactionnel, il existe un nombre de limites intrinsèques à cette approche impliquant l'acceptation des biais ainsi introduits. Pour exemple, seul le premier auteur est pris en compte, de plus il faut considérer les problèmes d'homonymie ou de fautes de frappe présentes

dans les bases de données. Les domaines sont inégalement représentés et les indicateurs bibliométriques s'appliquent très difficilement pour les sciences humaines et sociales. Toutes les revues ne sont pas recensées et pour celles qui le sont, il peut y avoir sur- ou sous-estimation de la revue et donc des travaux et des équipes. On notera que l'autocitation ou la citation d'un article controversé n'est pas abordée par l'approche statistique. De plus, les ouvrages ne sont pas pris en compte. Nous pouvons aussi constater que deux ans ne suffisent pas pour qu'un article se révèle or il s'agit de la durée retenue pour le calcul du facteur d'impact. Enfin, la citation négative n'est pas prise en compte. Pour le moment, il n'y a guère de solutions innovantes, seulement de nouvelles approches statistiques permettant de minimiser les biais introduits.

Méthodologie

Face à ce constat, il serait intéressant pour la communauté scientifique de disposer d'un outil plus qualitatif pour la conception de réseaux d'auteurs. Les outils de cartographie actuels s'appuient sur une approche quantitative et matricielle. Une nouvelle approche de cette problématique doit être envisagée. Sans prétendre fournir un traitement sémantique complet d'un article scientifique, nous pourrions dans un premier temps considérer les relations sémantiques entre l'auteur, les co-auteurs et les références bibliographiques. Il serait tout à fait pertinent de savoir si un article est cité de façon positive ou négative. Une référence bibliographique citée en contre-exemple est tout à fait révélatrice des relations entre les travaux des chercheurs. Il peut s'agir entre autres d'une référence par rapport à une définition, une hypothèse ou bien une méthode, mais également d'un point de vue, d'une comparaison ou bien d'une appréciation. Suite à l'identification des appels bibliographiques, nous proposerons une annotation de ceux-ci avec une catégorie afin de définir comment l'auteur a été cité. Cette catégorisation est définie par l'étude d'indices que nous relèverons dans la phrase. Nous rechercherons les indices positifs/négatifs de citation d'un auteur, ainsi que les citations hypothèses/méthodes utilisées par un auteur. On caractérisera alors ce point de vue comme étant une catégorisation sémantique des références de citation d'auteur. Le renvoi bibliographique qui se trouve dans le texte permet de définir un segment textuel où se trouvera l'information de catégorisation de ce renvoi. L'implémentation informatique de cette approche utilise la plateforme EXCOM (Exploration Contextuelle Multilingue) développée au sein du laboratoire LaLICC. Nous pourrions nous référer à l'article de [DJI06] décrivant plus en détail la plateforme.

Bibliographie et renvois bibliographiques

Selon Malcles, bibliographe, il est vrai que l'étude bibliométrique passe par une analyse de la bibliographie, mais Palanco souligna que l'application des statistiques à la bibliographie est réducteur, évoquant l'idée de réductionnisme bibliométrique puisque cette approche élimine la diversité des thèmes au profit de l'unité des matières. Nous prendrons comme postulat de départ que la bibliographie est effectivement une donnée essentielle pour l'évaluation des publications. L'appel de citation dans un texte peut prendre différentes formes. Il peut s'agir principalement

d'un renvoi numérique ou d'un renvoi par nom d'auteur. Pour cela, nous dresserons une classification des différentes familles numériques et alphanumériques des références bibliographiques.

Pour ce travail, nous avons utilisé les normes, mais également les « coutumes ». En effet, les renvois bibliographiques dans le texte sont plus ou moins normalisés selon les normes ISO 690-1 (Z 44-005) et ISO 690-2, mais il était nécessaire de prendre en compte des pratiques dépassant le simple renvoi numérique ou alphanumérique afin de pouvoir traiter exhaustivement l'ensemble des renvois bibliographiques. Afin de traiter automatiquement cette tâche d'identification et d'extraction, nous pourrions par exemple définir un alphabet adéquat permettant d'appliquer au corpus un automate fini déterministe. Pour identifier les renvois bibliographiques se trouvant présents dans le texte, nous nous appuyerons sur les travaux déjà effectués [BER06], qui proposent un automate à états finis afin de localiser les renvois bibliographiques. Cependant, au lieu de considérer l'aspect numérique d'une référence bibliographique, nous utilisons les renvois dans le texte afin de catégoriser les relations entre auteurs. Nous avons émis l'hypothèse que la pensée de l'auteur par rapport aux travaux de ses confrères se trouve à proximité de la référence bibliographique. Aussi considérons-nous dans cette première approche que la prise de position d'un auteur vis-à-vis de ces confrères se trouve dans un espace proche d'un renvoi bibliographique.

Indicateurs et indices

Nous nous proposons donc d'utiliser les renvois bibliographiques identifiés par l'automate à états finis d'un article afin de déterminer des segments textuels et déterminer un espace recherche sémantique associé à cette référence. Les renvois bibliographiques seront alors considérés comme étant nos indicateurs. Les indices linguistiques, quant à eux, permettent de déterminer une information sémantique spécifique. Ils permettent de réduire l'indétermination et de spécifier la qualité du renvoi. Il s'agit du seul savoir dont nous avons besoin pour déterminer nos catégories et se trouvent présents autour de l'indicateur, dans le même segment textuel que celui-ci. La méthode de l'Exploration Contextuelle, développée par Mr Desclés [DES91, DES96], va permettre à l'aide des indices, de lever les indéterminations sémantiques de l'unité linguistique analysée et proposer une catégorisation qualitative des références bibliographiques.

Segments textuels et localisation

L'indicateur permet de déterminer le segment textuel nécessaire et suffisant à l'accomplissement de notre tâche. Dans cette étude, nous ferons coïncider ce segment textuel avec la phrase. Nous nous gardons la possibilité d'étendre nos recherches à des zones plus larges, comme la théorie nous le permet, si cela s'avérait nécessaire à lever certaines ambiguïtés. Une fois l'espace de recherche déterminé, il faut prendre en compte la localisation de l'indice par rapport à l'indicateur. Nous avons identifié cinq localisations possibles par rapport à l'indicateur :

« **premier mot du segment textuel | avant le milieu | au milieu | après le milieu | à la fin du segment textuel** ». D'un point de vue pratique, seul le contexte **droit gauche** est implémenté et se révèle pour le moment suffisant dans le cadre de ce travail.

Catégorisation

Les différentes catégories ont été identifiées par Krushkov Yordan [KRU05] dans son travail de mémoire de maîtrise sous la direction de Mr Desclés. Elles se trouvent à la base de ce travail, aussi allons nous détailler les différentes catégories sur lesquelles nous nous appuyons.

Le point de vue est la première catégorie que nous avons identifiée. Il est très présent dans les corpus étudiés. Les indices linguistiques suivants font partie de cette catégorie :

« **Selon | d'après | pour | considérer que | nous y voyons | comme le dit |...** ».
Ils sont généralement localisés en amont de l'indicateur.

La seconde catégorie à laquelle nous nous sommes intéressés est la comparaison. En effet, nous comparons souvent le travail de nos confrères. Dans ce cas précisément, nous pouvons trouver des similarités ou bien des dissimilarités :

« **ressembler | comme dans les travaux de | le rapport avec |...** ».

Pour la non-ressemblance, nous avons comme indices linguistiques :

« **différer de | contraire l'approche de | contrairement ce qu'affirme |...** »

La catégorie de l'information est vaste. Pour cela, elle est divisée en sous-catégories comme l'hypothèse, l'analyse et le résultat. Pour la sous-catégorie de l'analyse, nous pouvons donner comme exemple :

« **a été analysé dans | l'analyse de | lors de son analyse | ...** ».

Pour concevoir la sous-catégorie des résultats, nous avons considéré les indices linguistiques suivants :

« **nous avons démontré | donner de nombreux exemples de | a publié ses résultats | a dégagé |...** »

La catégorie de la définition est également importante avec pour indices :

« **ils caractérisent | la notion ... introduite dans |...** »

La catégorie de l'appréciation met en valeur le jugement d'un auteur sur un autre auteur ou plutôt sur un ou plusieurs travaux de celui-ci. Il peut s'agir d'un jugement positif ou négatif :

« **ont rejeté | n'as pas répondu | en trahissant sérieusement notre proposition | ...** ».

Cette catégorie est très importante dans le sens où elle apporte une réponse à l'un des biais introduits par l'approche statistique.

Quotation	Point de vue Soi-même Autrui	Pris de position
	Comparaison Soi-même Autrui	Similitude
		Dissimilitude
	Information Soi-même Autrui	Hypothèse
		Analyse
		Résultat
		Méthode
		Citation
	Contre-exemple	
Definition Soi-même Autrui		
Appréciation Autrui	Accord	
	Désaccord	

Figure 1 : Catégorisation des renvois bibliographiques

Constitution du corpus

Pour cette étude, nous avons constitué un corpus d'articles issus du laboratoire LaLICC afin d'identifier les indices et de constituer notre base de connaissances. Afin de traiter le caractère pluridisciplinaire de notre approche, nous avons augmenté le corpus avec des publications extraites de HAL, la base de données de l'INRIA. Nous avons également choisi des articles de la revue INTELLECTICA. Ce petit corpus de test couvre les domaines de la linguistique, de l'informatique, et des sciences cognitives afin de démontrer la capacité du système à traiter une information multidisciplinaire. À la rédaction de cet article, le corpus est exclusivement constitué de textes en français. La couverture de l'anglais sera une prochaine étape dans le développement de ce système.

Plateforme informatique

L'architecture informatique de la machine à annoter automatiquement EXCOM, qui s'inspire de l'architecture modulaire GATE, est décrite dans la figure suivante. Les textes traités par EXCOM sont d'abord prétraités pour les préparer à une segmentation en phrases, paragraphes et sections en s'appuyant sur les travaux de [MOU99a, MOU99b].

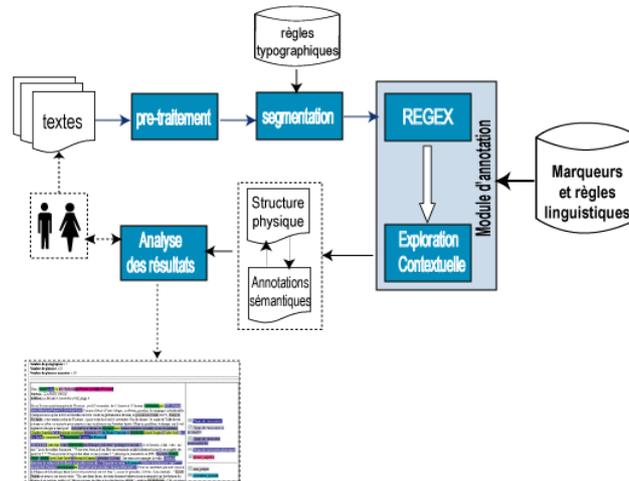


Figure 2 : Architecture informatique de la machine EXCOM

À chaque tâche d'annotation sémantique sont associés un ensemble de marqueurs linguistiques (listes d'indicateurs et d'indices) et un ensemble de règles applicables sont les textes segmentés. Les conditions de déclenchement de ces règles sont exprimées de différentes façons qui déclenchent certains niveaux du moteur d'annotation. Chaque niveau fait appel à un algorithme général de fonctionnement. Ce moteur est construit sous une forme multicouche où chaque brique répond à un besoin d'annotation particulier. Les briques les plus élémentaires sont définies pour répondre à un besoin d'annotation des références bibliographiques (indicateurs du point de vue bibliosémantique) sous forme d'expressions régulières et les plus externes prennent en charge les aspects d'une exploration contextuelle. Les modules les plus externes s'appuient sur les modules les plus internes.

Le module REGEX fait appel à un moteur d'expressions régulières. Les domaines d'utilisation des expressions régulières sont nombreux : elles interviennent dans le cadre de l'analyse de contenu des textes. Avec le support d'Unicode, l'extraction d'information peut se réaliser sur des documents multilingues.

Le module d'exploration contextuelle (EC) est composé :

- d'un ensemble de marqueurs linguistiques (indicateurs et indices) ;
- d'un ensemble de règles d'EC qui se présentent sous la forme de règles déclaratives (si certaines conditions sont vérifiées alors certaines actions sont appliquées).
- d'un moteur d'EC qui applique les règles en respectant la primauté de l'indicateur sur les indices complémentaires.

Le résultat de l'application de ces règles est un texte annoté. Les annotations sont des marques sous forme d'éléments et attributs XML. La sémantique de ces annotations est liée à l'organisation de la catégorie du point de vue reconnue par le système EXCOM. L'objectif de cette plateforme est de proposer une exploration du texte afin de l'augmenter d'informations sémantiques sous forme d'annotations. Si la plupart des travaux menés dans ce domaine s'appuient sur une analyse morpho-syntaxique, la méthode préconisée pour cette plateforme est l'Exploration Contextuelle et utilise une base de connaissances, constituée de marqueurs linguistiques. Elle permet d'étiqueter automatiquement un texte à partir de ressource linguistique.

Déclaration de Règles

L'application informatique nécessite l'écriture de règles. Celles-ci se présentent sous la forme d'un fichier XML. Aussi allons-nous détailler une règle qui permet d'annoter la publication selon le point de vue de la *méthode*, qui est une sous-catégorie de *information*.

```
<regle          nom_regle="ReglInfMet3"          tache="bibliosemantique"
point_de_vue="information" type="EC">
<conditions>
  <indicateur  espace_de_recherche="phrase"    type="annotation"
    valeur="RenvBiblio" />
  <indice     contexte="droite"    espace_de_recherche="."    type="liste"
    valeur="IdAuteur" />
  <indice     contexte="droite"    espace_de_recherche="."    type="liste"
    valeur="IdMethode" />
</conditions>
<actions>
  <annotation type="ajout_attribut" espace="identique" annotation="methode" />
</actions>
</regle>
```

Cette règle traite donc du point de vue de l'information : *point_de_vue="information"*. L'indicateur a pour valeur : *valeur="RenvBiblio"* qui permet de retrouver les renvois bibliographiques et identifier l'espace de recherche qui est la phrase : *espace_de_recherche="phrase"*. Les indices, de type liste, sont définis par leur contexte qui peut être **droite** ou **gauche** par rapport à l'indicateur, dans l'espace de recherche préalablement identifié. Dans le cas présent, les deux indices se trouvent à droite de l'indicateur. Si l'ensemble des conditions de cette règle est validé, alors EXCOM annote le segment textuel en ajoutant un attribut : *<annotation type="ajout_attribut" espace="identique" annotation="methode" />*

Résultats

Les résultats sont affichés sous la forme suivante : Le segment textuel est coloré en bleu. L'indicateur est en vert et les indices primaires et secondaires sont respectivement en vert clair et mauve.

de la largeur des données dans la première phase est réalisée sous contrainte de précision, mais la seconde phase affecte des opérations sur des opérateurs de largeur plus importante qui se traduit par une augmentation de la précision des calculs. En conséquence, la solution obtenue n'est pas optimisée exactement pour la contrainte de précision spécifiée. Dans [10], les auteurs proposent une méthode pour laquelle, la synthèse d'architecture est effectuée entre deux phases d'optimisation de la largeur des données. Le partage de ressources est pris en compte pour réduire à la fois le coût matériel et le temps d'optimisation. En effet, l'évaluation de la précision étant réalisée par simulation il est nécessaire de réduire l'espace de recherche par l'utilisation d'heuristiques afin d'avoir des temps d'optimisation raisonnables. Une première étape analyse le graphe flot de signal de l'application pour former des groupes de données. Une même largeur sera assignée aux données d'un même groupe. Ainsi, par exemple, les entrées d'additions successives pourront former un seul groupe. La seconde étape détermine la largeur de données minimale requise pour chaque groupe. Enfin, cette méthode permet de fixer indépendamment et simultanément le coefficient de précision des données primaires et secondaires.

Figure 3 : Exemple du point de vue méthode

Discussion

Le premier point que nous discuterons est celui des renvois bibliographiques. L'étude des segments textuels repose principalement sur l'identification de ces renvois. Aussi est-il très important dans cette approche que l'ensemble des renvois soit reconnu. Si sur cet exemple, aucun problème d'identification n'a mis à défaut cette approche, il faudra cependant tenir compte, sur des corpus plus littéraires, de la notion de courant ou de personnes associées en tant que telles. Par exemple, « *Selon Pottier, nous devons concevoir que ...* ».

Le deuxième point est la nécessité de continuer le travail linguistique afin de pouvoir couvrir l'ensemble des catégories identifiées et de les implémenter au sein de la plateforme. Seule la volumétrie nous permettra de proposer à cette approche qualitative un protocole dévaluation.

Le troisième point est une remarque quantitative. Sur cet exemple, nous avons constaté que l'auteur se référait plusieurs fois à la même publication d'un de ces confrères selon le point de vue de la méthode. L'identification des renvois bibliographiques peut donc apporter une pondération à l'outil bibliométrique. Cependant, il faut bien souligner que notre approche, va au-delà d'une simple pondération puisqu'à une référence bibliographique, nous faisons correspondre une catégorisation sémantique.

Conclusion

À court terme, cette approche permettra de proposer un outil beaucoup plus fin et complémentaire à l'approche proposée actuellement. D'une part, la prise en compte de la bibliographie comme unité est loin d'être satisfaisante et de nombreux biais sont introduits. Le fait de pouvoir catégoriser la bibliographie par une analyse linguistique donc qualitative et automatisé via la plateforme informatique EXCOM offrira un outil plus pertinent et offrira à moyen terme de nouvelles possibilités

d'exploration des textes scientifiques. D'autre part, cette approche permet de catégoriser sémantiquement les renvois bibliographiques. Aussi et contrairement à une approche statistique, pouvons-nous étudier et obtenir des résultats sur un très petit nombre de publications, à l'échelle d'un laboratoire par exemple, tout en conservant la possibilité de travailler à une plus grande échelle. À l'approche statistique de l'évaluation, l'approche linguistique permet de porter un regard qualitatif des relations entre les travaux des différents auteurs.

Perspectives

Cette étude servira de point de départ à une nouvelle façon de cartographier la science ou du moins un domaine. L'utilisation d'un logiciel comme Pajek [BAT01], [BAT02] et [BAT05] va nous permettre d'analyser des réseaux sous forme de graphes en permettant d'annoter les arcs non pas une pondération mais une catégorie sémantique. Nous serons alors à même de déterminer des ensembles de sous-graphes en fonction des catégories précédemment établies.

Bibliographie

[BAT01] Batagelj, V., Pajek - program for large networks analysis and visualization Presented at Dagstuhl seminar Link Analysis and Visualization Dagstuhl 1-6. July 2001.

[BAT02] Batagelj V., Mrvar A., Zaveršnik M., "Network Analysis of Texts". Jezikovne tehnologije / Language Technologies, T. Erjavec, J. Gros eds., Ljubljana 2002, p. 143-148.

[BAT05] Batagelj, V. Brandes U., "Efficient generation of large random networks", 2005 Physical Review E 71, 036113.

[BER06] Bertin M., Desclés J.P., Djioua B., Krushkov Y., "Automatic Annotation in Text for Bibliometrics Use", FLAIRS 2006, Floride, 11-13 mai

[BON82] Bonzi, S., "Characteristics of a literature as predictors of relatedness between cited and citing works". Journal of the American Society for Information Science, 1982, 33(4): 208-216

[BRA03] Bradshaw S., "Reference directed indexing: Redeeming relevance for subject search in citation indexes". In Proceedings of the 7th European Conference on Research and Advanced Technology for Digital Libraries, 2003.

[COL92] Cole, S., "Making Science. Between Nature and Society", 1992, Harvard University Press, Cambridge, MA.

[CRO84] Cronin, B., "The Citation Process: The Role and Significance of Citations in Scientific Communication", 1984, Taylor Graham, London.

[DES91] Desclés, J. P., "Exploration contextuelle et sémantique: un système expert qui trouve les valeurs sémantiques des temps de l'indicatif dans un texte.", 1991, Knowledge modeling and expertise transfert p.371–400.

[DES97] Desclés, J. P., "Système d'exploration contextuelle." Co-texte et calcul du sens p.215–232. 1997.

[DJI06] Brahim, D. , Flores, J.G., Blais, A., Desclés J-P., Gael, G., Jackiewicz, A., Le Priol, F., Leila,N.B., Sauzay B., "EXCOM: an automatic annotation engine for semantic information", FLAIRS 2006, Floride, 11-13 mai,

[GAR65] Garfield, E., "Can citation indexing be automated ?" National Bureau of Standards Miscellaneous.1965. Publication, 269:189–192

[KIN87] King, J., "A review of bibliometric and other science indicators and their role in research evaluation", Journal of Information Science, 1987. Vol. 13 No. 5, pp. 261-76.

[KRU05] Krushkov, Y. (2004-2005), "L'exploration contextuelle des appariements entre les références bibliographiques et les passages textuels dans un corpus de textes linguistiques." Mémoire de Maîtrise, Université Paris IV Sorbonne sous la dir. de Mr J.P Desclés.

[LAT87] Latour, B., Science in Action, Open University, Milton Keynes. 1987.

[LEY98] Leydesdorff, L., "Theories of citation ?", Scientometrics, 1988, Vol. 43 No. 1, pp. 5-25.

[LIP65] Lipetz, B. A., Improvements of the selectivity of citation indexes to science literature through inclusion of citation relationship indicators. American. Documentation, 16:81–90, 1965.

[LIU93] Liu, M., "Progress in documentation: the complexities of citation practice - a review of citation studies", Journal of Documentation, 1993. Vol. 49 No. 4, pp. 370-408.

[MER04, MAR04] Mercer, R. E. and Marco, C. D., "A design methodology for a biomedical literature indexing tool using the rhetoric of science". In BioLink workshop in conjunction with NAACL/HLT, pages 77–84. 2004.

[MOR75, MUR75] M. J. Moravcsik and P. Murugesan., "Some results on the function and quality of citations." Social Studies of Science, 5:86–92.1975

[MOU99a] Mourad Ghassan, "Rôle de la typographie dans la segmentation de textes", JILA'99 (Journées Internationales de Linguistique Appliquée), p.203-206.

[MOU99b] Mourad Ghassan, "La segmentation de textes par l'étude de la

ponctuation", CIDE'99 (2e Colloque International sur le Document Électronique), p.155-171.

[NAN00 et al.] Nanba, H. Kando, N. and Okumura, M., "Classification of research papers using citation links and citation types: Towards automatic review article generation". In American Society for Information Science SIG Classification Research Workshop: Classification for User Support and Learning, pages 117–134, 2000.

[PRI63] Price, D., "Little Science, Big Science", p.IV-V. 1963.

[PRI65] Price, D. De Solla, "Networks of scientific papers", Science, 1965 Vol. 149, pp. 510-5.

[PRI86] Price, D. De Solla, "Little Science, Big Science . . . And Beyond", Columbia University Press, New York, NY.1986

[SCH04] Schneider J.W., "Introduction to bibliometrics for construction and maintenance of thesauri". Journal of documentation. 2004. Vol.60 N°5 p.524-549

[TEU00, MOE00] TEUFEL,S. MOENS, M., "What's yours and what's mine: Determining Intellectual Attribution in Scientific Text" In: Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. Hong Kong, Oct 2000.

[TEU01] TEUFEL S., "Task-Based Evaluation of Summary Quality: Describing Relationships Between Scientific Papers". Workshop Automatic Summarization', NAACL-2001.

[WHI04] White H. D., "Citation analysis and discourse analysis revisited". Applied Linguistics, 25(1):89–116. 2004.

[WIL99] Wilson, C.S., "Informetrics", in Williams, M.E. (Ed.), Annual Review of Information Science and Technology, Vol. 34, pp. 107-247.1999