De la recherche d'informations hautement spécialisées : le cas de la recherche d'informations dans les brevets de chimie

Patrick Ruch
patrick.ruch@hesge.ch
https://orcid.org/0000-0002-3374-2962
Haute Ecole de Gestion, Genève

Résumé

Nous décrivons le développement d'un moteur de recherche avancé pour la recherche d'informations dans les bibliothèques de brevets de chimie. Nous utilisons la campagne internationale d'évaluation TREC (Text Retrieval Conferences) pour évaluer une stratégie de recherche combinant : un modèle de recherche vectoriel standard, les réseaux de co-citations reliant les brevets, et une stratégie de normalisation (synonymes ramenés à un identifiant unique) des entités nommées chimiques basée sur le traitement automatique de la langue. Un moteur vectoriel basique obtient une précision moyenne de 0.067. On observe qu'un gain de précision important est apporté par l'usage des réseaux de citations (+168%), tandis que d'autres contenus, tels les codes IPC, semblent n'apporter aucun gain. Conclusion : Les performances de notre moteur (précision moyenne proche de 20%), développé en quelques semaines seulement, le placent en tête des évaluations officielles TREC ; ce qui suggère que la valeur d'une collection porte davantage sur son contenu que sur les instruments de recherche, désormais à la portée de n'importe quelle équipe de développeurs en science de l'information.

Abstract

We describe the development of an advanced retrieval engine to search information in libraries of chemical patents. We use the international TREC (Text Retrieval Conferences) evaluation framework to develop and assess an original search strategy combining a standard vector-space model, a network of co-citations between patents, and a strategy of standardization/expansion of chemical named-entities based on natural language processing. Our basic engine obtains an average accuracy of 0.067. The most significant precision gain is provided by the use of co-citations (+168%), while other contents, in particular ICP codes, do not improve retrieval effectiveness of the engine. The official TREC performance of our engine (ranked #1, with a mean average precision approaching 20%) emphasize the role of document contents as opposed to technological expertise in document retrieval.

Mots-clés

modèles de recherche d'information, bibliothèque numérique, propriété intellectuelle, chimie, indexation models of information retrieval, digital library, IP, chemistry, indexing

1. Introduction

Pour la plupart d'entre nous, professionnels de l'information ou citoyens profanes de la société éponyme, la notion de recherche d'information évoque spontanément l'usage de moteurs de recherche portant sur des contenus du web, que nous utilisons quotidiennement pour de nombreuses tâches, dont certaines sont en effet assimilables à de la recherche d'information, la recherche d'information dans des données financières (*Yahoo! Finance*) ou dans une collection de messages électroniques.

S'il est difficile de caractériser précisément ces différents usages, tant il est désormais possible de mélanger d'une manière transparente pour l'utilisateur l'ensemble des flux de recherche, on peut toutefois observer qu'un certain nombre de contenus demeure plus difficilement accessible via ces outils. Parmi ces contenus, les scientifiques de nos Hautes Ecoles ont remarqué à quel point les bibliothèques scientifiques, dont certaines sont pourtant indexés par ces outils, sont mal représentées dans les résultats retournés par ces mêmes outils. C'est là un paradoxe qui mérite d'être remarqué et qui est bien documenté en tout cas pour le domaine biomédical. En effet, la charge d'un moteur de recherche comme PubMed et des autres moteurs et ressources mis à la disposition par l'institution hôte de PubMed, le NCBI, ne fait que croître avec plus d'un million d'accès quotidiens. Encore plus remarquable, l'arrivée de nombreuses alternatives à PubMed, telles que la GlobalHealthLibrary (http://www.globalhealthlibrary.net/php/index.php), supportée par l'OMS, SRS (Sequence Research System ; http://srs.ebi.ac.uk/), maintenu par l'EBI (European Bioinformatics Institute), ou EAGLi (Engine for Answering questions in Genomic Literature : http://eagl.unige.ch/EAGLi/), développé à Genève, avec des spécialisations toujours plus poussées (e.g. génomique), suggère que l'accès réel à MEDLINE via des moteurs très pointus est largement sous-estimé par les statistiques d'accès au seul moteur PubMed.

Le succès des moteurs dédiés à MEDLINE peut s'expliquer pour différentes raisons. D'abord, ces outils spécialisés fournissent au chercheur des méthodes de recherches adaptées spécifiquement à l'information disponible dans la bibliothèque médicale, ce qui permet d'optimiser la pertinence des recherches par rapport à un moteur généraliste, plutôt adapté à une topologie documentaire du type web et aux protocoles basés sur les hyperliens (Brin and Page 1998). Ensuite, ces outils spécialisés ont innové en mettant à la disposition du chercheur un ensemble de ressources de navigation spécifiques, accessibles via un treillis de liens croisés vers un certain nombre de bases de connaissances du domaine. C'est notamment le cas de PubMed, qui n'est qu'une source parmi des douzaines de bases de données, bibliothèques, et vocabulaires contrôlés, mis à disposition du public par le NCBI (National Center for Biotechnological Information : http://www.ncbi.nlm.nih.gov/), directement rattaché à la NLM (National Library of Medicine) du NIH (National Institute of Health). On retrouve ce modèle avec EBIMed et EAGLi qui peuvent renvoyer directement l'utilisateur vers des bases de connaissances telles qu'UniProt (Bairoch et al. 2005) en fonction du contenu informationel proposé par ces moteurs.

Enfin, le succès de ces moteurs hautement spécialisés s'appuie sur le dynamisme d'une communauté de chercheurs, qui a parfaitement suivi, et bien souvent conduit les développements les plus avancés des sciences de l'information, avec l'avènement d'un champ spécialisé appelée bioinformatique. La bioinformatique suit maintenant une voie qui lui est propre, avec une richesse conceptuelle et une productivité opératoire et scientifique

proprement vertigineuse aboutissant à des « ruptures » épistémologiques d'une fréquence sans précédent dans l'histoire des sciences et des techniques. On peut citer l'avènement de « nouvelles » sciences telles que la génomique, qui porte sur les gènes ; de la protéomique, qui porte sur les protéines ; de la métabolomique, qui porte sur des molécules plus petites que les gènes et produits de gènes ; et enfin, de la bibliomique, qui utilise la littérature pour étudier les autres « omiques » (Grivell 2002) !

S'il est permis de douter que toutes ces spécialités et domaines de recherche pourront acquérir la stabilité nécessaire à leur transformation en *sciences normales* autonomes, force est de constater que le modèle historique du développement scientifique, basé sur la remise en cause radicale de *paradigmes* scientifiques, tels que définis par Kuhn (1962) semble remis en cause par la rapidité à laquelle biologie moléculaire et informatique synthétisent de nouveaux champs de recherche.

Dans ce contexte, la littérature scientifique joue clairement un rôle central avec une activité importante, comme en témoigne le million de requêtes quotidiennes reçues par PubMed. Relativement plus discrètes et comparativement moins visibles, bien que tout aussi importante, les bibliothèques de brevets et leurs moteurs de recherche (e.g. esp@cenet) ont reçu récemment une attention particulière. En effet, celles-ci ont la particularité d'être à la fois relativement universelles dans leurs couvertures et hautement spécialisées dans leurs contenus. Alors que la lecture d'un brevet en chimie ou en électronique demande des connaissances pointues, on observe que les grandes bases de données de brevets (e.g. EPO, European Patent Office : http://www.espacenet.com/access/index.en.htm ou USPTO, US Patent and Trademark Office : http://www.uspto.gov/patents/process/search/index.jsp) couvrent à peu près la totalité du spectre techno-scientifique, tel que représenté par la classification internationale des brevets, qui totalise environ 70'000 descripteurs. Cette spécialisation, tant au niveau des bibliothèques dans le domaine des sciences de la vie que des bases de données de brevets, constitue une singularité remarquable vis-à-vis de cette autre grande source d'information documentaire que constitue le web.

Nous pensons en particulier que dans le type de contenus que sont les brevets, la spécialisation requise est de nature à limiter l'hégémonie observée dans le monde des moteurs de recherche du web. Remarquons que l'expertise méthodologique à mettre en œuvre appartient au corpus classique de la science de l'information (Capurro & Hjorland 2002) autour de concepts théoriques liés à recherche d'information (indexation, méta-données, expansion de requête, jugements de pertinence…), à la bibliométrie (citations, facteurs d'impact...), et à l'usage des langages documentaires (descripteurs, ontologies…). Cette expertise, qui constitue l'un des piliers de la formation dans la filière information-documentaire, notamment dans sa spécialisation en systèmes d'informations, suffit au développement d'outils avancés susceptibles de rivaliser avec des outils commerciaux. Dans le cadre d'une procédure expérimentale très contrôlée et dédiée à une tâche particulière, à savoir une tâche de génération automatique de la bibliographie décrivant l'état de l'art d'une invention – aussi appelé recherche d'antériorité – nous verrons comment une combinaison holistique de ces différentes composantes (indexation, méta-données, citations…) permet de développer un moteur de recherche original et performant.

L'exposé se déclinera comme suit : tout d'abord, nous présentons brièvement les Text Retrieval Conferences, qui servent de cadre à l'étude que nous avons réalisée ; ensuite, nous décrivons

N° DOI : 10.55790/journals/ressi.2010.e1102

les données utilisées pour nos expériences ; puis nous détaillons la méthodologie de développement du moteur de recherche ; enfin nous récapitulons les résultats obtenus dans le cadre de la campagne d'évaluation TREC que nous discutons avant de conclure notre article.

2. Text Retrieval Conferences

Partant du constat que ni les moteurs de recherche d'informations dans MEDLINE, ni les outils d'interrogation de bibliothèques de brevets n'avaient été affectés par l'avènement des moteurs de recherche du web, les organisateurs des TRECs (Text Retrieval Conferences) décidèrent dès 2002 de s'intéresser à cette question.

2.1. Un peu d'histoire

Chaque année depuis plus de vingt ans, les agences gouvernementales NIST (National Institute of Standards and Technology) et DARPA (Defense Advanced Research Projects Agency) organisent les Text Retrieval Conferences (TREC, http://trec.nist.gov/), une campagne d'évaluation pour mesurer les avancées dans le domaine de la recherche d'informations. Les laboratoires des universités les plus prestigieuses, ainsi que les centres de recherches des sociétés les plus actives du web (Microsoft, Google, Yahoo...) participent régulièrement à ces compétitions, que les agences nord-américaines de financement de la recherche suivent également avec attention.

Héritières du *Cranfield paradigm* (cf. Capurro et Hjorland 2002), les campagnes d'évaluations TREC reposent sur le développement d'une collection (*benchmark*) contenant trois composantes : un ensemble de requêtes ; un corpus de documents ; des jugements de pertinence associant chaque requête avec un ou plusieurs documents jugés « pertinents ».

2.2. Recherche documentaire spécialisée

Ainsi, entre 2002 et 2007, diverses tâches de recherche d'informations dans la bibliothèque numérique MEDLINE furent-elles proposées et évaluées dans le contexte de la « Genomics track ». Plus récemment, en 2009, une tâche de recherche d'information dans une base documentaire de brevets de chimie fut proposée (« Chemical patent retrieval track »). Dans ce dernier cas, il s'agissait de modéliser une tâche d'analyse de l'état de l'art (*prior art search*), telle que pratiquée communément par les déposants de brevets, ainsi que par les experts des bureaux de brevets, employés à l'EPO (European Patent Office), USPTO (United States Patent and Trademark Office), ou plus proche de nous à Berne à l'IFPI (Institut Fédéral de la Propriété Intellectuelle). Pour la première année, la tâche attirait une quinzaine d'institutions privées et publiques à la distribution géographique très variée dont Carnegie Mellon University, Dalian University of Technology, Erasmus University, Fondazione Ugo Bordoni, Fraunhofer Institute, Haute Ecole de Gestion, MerckSerono, Milwaukee School of Engineering, Purdue University, Salalah College of Technology, TGN Corporation, University of Alcala, University of Iowa, York University. Une présentation détaillée des procédures d'évaluations, des métriques, et des résultats officiels se trouve dans Lupu et al., 2010.

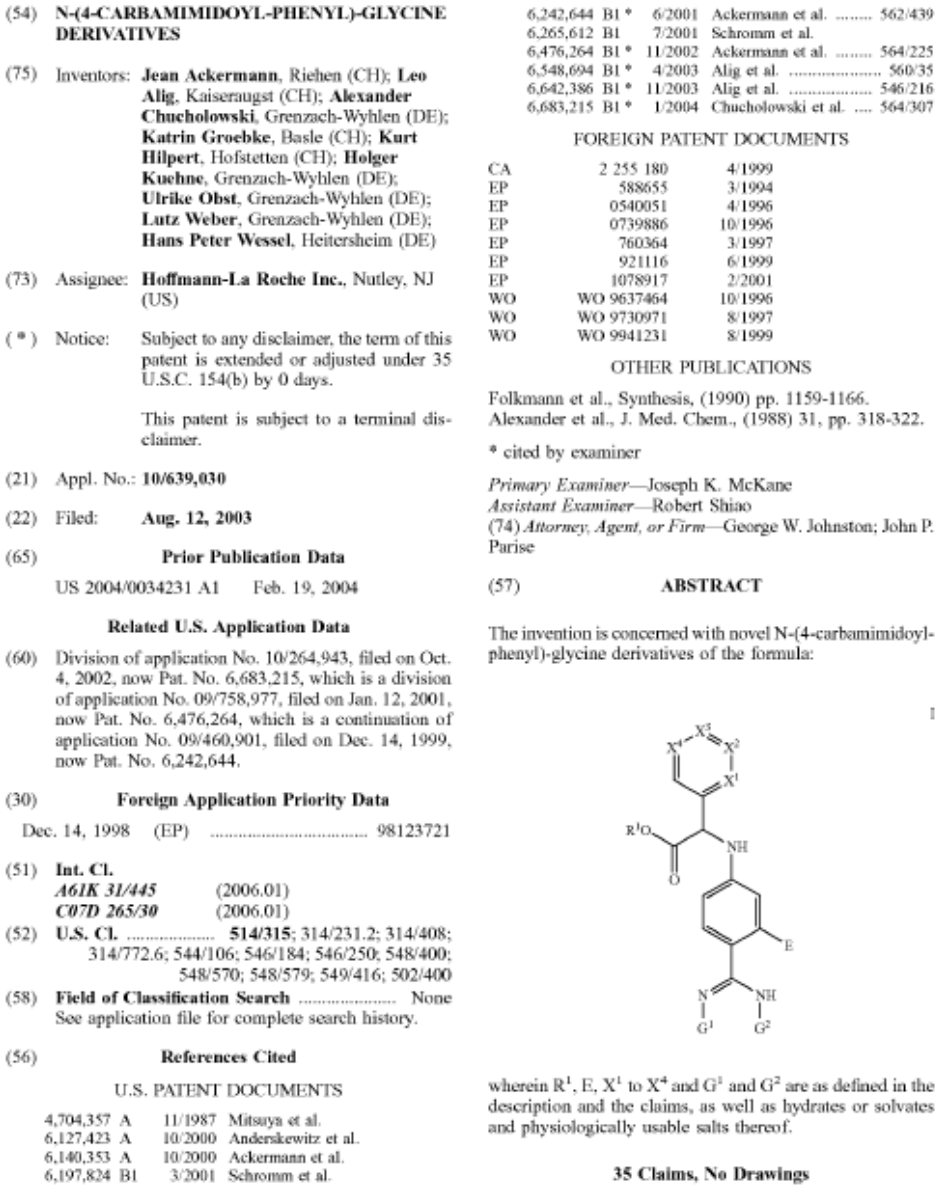
3. Données

Une collection de 1.2 million de brevets au format XML est fournie par les organisateurs. Le total représente un peu moins de 2 millions de documents, incluant les différentes versions d'un brevet (version initiale ou après examen) et les fichiers associés, dont des images

N° DOI : 10.55790/journals/ressi.2010.e1102

(tableau, structures chimiques, images…) et des séquences de nucléotides. Avec ce montage expérimental, les jugements de pertinence, qui relient les requêtes aux documents de la collection, sont directement disponibles sans qu'il soit nécessaire de procéder à des évaluations par des experts, lesquelles sont coûteuses et difficiles à mettre en place. Un exemple de document contenant certains des champs les plus significatifs d'un brevet est fourni en Figure 1. Nos réglages préliminaires suggéraient que certains champs présentaient un intérêt particulier pour la recherche d'informations, notamment le Titre, la Description (le champ le plus long) ; le Résumé, les « Claims » (ou *Revendications*), les codes IPC (International Patent Classification). Un exemple de références bibliographiques, utilisées comme jugements de pertinence pour l'évaluation, se trouve également en Figure 1 (section 56 : « References Cited »). Les champs *Inventeur* et *Applicant* ne furent arbitrairement pas retenus. Nous verrons dans la discussion que certains de ces choix méritent d'être questionnés a posteriori.

Figure 1 : Exemple de brevets avec certains des champs les plus significatifs.



(54) **N-(4-CARBAMIMIDOYL-PHENYL)-GLYCINE DERIVATIVES**

(75) Inventors: **Jean Ackermann**, Riehen (CH); **Leo Alig**, Kaiseraugst (CH); **Alexander Chucholowski**, Grenzach-Wyhlen (DE); **Katrin Groebke**, Basle (CH); **Kurt Hilpert**, Hofstetten (CH); **Holger Kuehne**, Grenzach-Wyhlen (DE); **Ulrike Obst**, Grenzach-Wyhlen (DE); **Lutz Weber**, Grenzach-Wyhlen (DE); **Hans Peter Wessel**, Heitersheim (DE)

(73) Assignee: **Hoffmann-La Roche Inc.**, Nutley, NJ (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

This patent is subject to a terminal disclaimer.

(21) Appl. No.: **10/639,030**

(22) Filed: **Aug. 12, 2003**

(65) **Prior Publication Data**

US 2004/0034231 A1 Feb. 19, 2004

Related U.S. Application Data

(60) Division of application No. 10/264,943, filed on Oct. 4, 2002, now Pat. No. 6,683,215, which is a division of application No. 09/758,977, filed on Jan. 12, 2001, now Pat. No. 6,476,264, which is a continuation of application No. 09/460,901, filed on Dec. 14, 1999, now Pat. No. 6,242,644.

(30) **Foreign Application Priority Data**

Dec. 14, 1998 (EP) 98123721

(51) **Int. Cl.**
A61K 31/445 (2006.01)
C07D 265/30 (2006.01)

(52) **U.S. Cl.** **514/315**; 314/231.2; 314/408; 314/772.6; 544/106; 546/184; 546/250; 548/400; 548/570; 548/579; 549/416; 502/400

(58) **Field of Classification Search** ..... None
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

4,704,357	A	11/1987	Mitsuya et al.
6,127,423	A	10/2000	Anderskewitz et al.
6,140,353	A	10/2000	Ackermann et al.
6,197,824	B1	3/2001	Schromm et al.
6,242,644	B1 *	6/2001	Ackermann et al. 562/439
6,265,612	B1	7/2001	Schromm et al.
6,476,264	B1 *	11/2002	Ackermann et al. 564/225
6,548,694	B1 *	4/2003	Alig et al. 560/35
6,642,386	B1 *	11/2003	Alig et al. 546/216
6,683,215	B1 *	1/2004	Chucholowski et al. 564/307

FOREIGN PATENT DOCUMENTS

CA	2 255 180	4/1999
EP	588655	3/1994
EP	0540051	4/1996
EP	0739886	10/1996
EP	760364	3/1997
EP	921116	6/1999
EP	1078917	2/2001
WO	WO 9637464	10/1996
WO	WO 9730971	8/1997
WO	WO 9941231	8/1999

OTHER PUBLICATIONS

Folkmann et al., Synthesis, (1990) pp. 1159-1166.
Alexander et al., J. Med. Chem., (1988) 31, pp. 318-322.

* cited by examiner

Primary Examiner—Joseph K. McKane
Assistant Examiner—Robert Shiao
(74) *Attorney, Agent, or Firm*—George W. Johnston; John P. Parise

(57) **ABSTRACT**

The invention is concerned with novel N-(4-carbamimidoyl-phenyl)-glycine derivatives of the formula:

I

wherein R^1, E, X^1 to X^4 and G^1 and G^2 are as defined in the description and the claims, as well as hydrates or solvates and physiologically usable salts thereof.

35 Claims, No Drawings

N° DOI : 10.55790/journals/ressi.2010.e1102

En plus des requêtes utilisées pour l'évaluation officielle, pour lesquels les jugements de pertinence étaient finalement mis à disposition des compétiteurs après l'évaluation officielle, une collection de cinq cents requêtes avec jugements de pertinences était fournie pour le développement des moteurs de recherche.

4. Méthodes

Dans cette section, nous présentons les principaux éléments de notre stratégie de recherche. Une présentation plus détaillée est disponible dans Gobeill et al. 2009 et Gobeill et al. 2010. Afin de développer la stratégie de recherche et régler la combinaison des différents modules, nous cherchons à maximiser une fonction objective : la précision moyenne. Notre système de base utilise un moteur de recherche vectoriel (Ruch 2006). Avec un tel outil et après le calcul des paramètres traditionnels tels que le choix du schéma de pondération et son réglage, nous obtenons une Précision Moyenne (PM) de 0.051.

4.1. Représentation des requêtes et des documents

En ce qui concerne la base documentaire, la première étape consiste à la condenser. En effet, pour des raisons d'efficacité (réduction de volume des index, temps de réponses…), il est fréquent qu'une partie seulement d'un document soit indexée. En particulier, la *Description* du brevet devait être réduite. Après différentes tentatives de condensation plus ou moins fructueuses sur critères statistiques et positionnels, il apparaît rapidement qu'un retrait pur et simple de cette section amène un gain de 30% (PM de 0.067). Un « raciniseur » de type Porter légèrement modifié (van Reijsbergen et. al. 1980), et une liste de mots-outils d'environ 500 mots sont utilisés en complément pour réduire la taille de l'index (Dolamic and Savoy 2010).

Plus systématiquement, nous évaluons ensuite la contribution individuelle de chacun des champs du brevet à la recherche de l'état de l'art. Cette contribution est évaluée relativement au système combinant tous les champs d'un brevet, à l'exception de la description, que nous utilisons comme modèle de *Référence*. Dans le Tableau 1, nous voyons l'impact du retrait d'une section particulière du brevet section par section.

Tableau 1 : Precision moyenne selon les champs retenus.

Référence/Champs retirés	MAP
Référence	0.067
Title	0.066
Abstract	0.063
Claims	0.036
IPC 4-digits codes	0.055
IPC complete codes	0.058

N° DOI : 10.55790/journals/ressi.2010.e1102

On constate que le retrait de la section "Claims", qui contient la liste des innovations apportées par l'invention, fait chuter la précision moyenne de près de 50%, de 0.067 à 0.036. Plus conforme à l'intuition (Tseng and Wu 2008), le retrait d'autres éléments du brevet comme le titre représente un impact clairement moindre. Le retrait du résumé constitue également une perte d'information relativement faible. En comparaison, le retrait des codes IPC affecte plus significativement les résultats. Symétriquement, au niveau des requêtes, la même sélection est appliquée et les mêmes champs sont utilisés, à l'exception notable de la description, qui est conservée pour les requêtes. En effet, nous observons que dans les requêtes, un gain de 3% est mesuré lorsque la description est présente sans que nous ne soyons en mesure pour l'instant d'expliquer une telle dissymétrie entre la requête et la collection.

4.2. Modèle de recherche

Les expériences sont effectuées avec un schéma de pondération probabiliste dont le réglage fin nous permet d'atteindre une PM de 0.073 ; cf. Gobeill et al. 2009 pour une présentation détaillée des réglages statistiques. En complément du résultat ordonné fourni par la recherche d'information probabiliste, nous explorons deux stratégies supplémentaires : le filtrage selon la classification internationale des brevets, et le reclassement selon le nombre de citations de chaque brevet.

En ce qui concerne l'utilisation des descripteurs IPC, certains spécialistes (Sternitzke, 2009) de l'analyse de l'état de l'art (*recherche d'antériorité*), prétendent qu'une recherche sur un code IPC à 4 valeurs permet de collecter la totalité de l'état de l'art. Plus modestement, d'autres (Criscuolo and Verspagen 2008), affirment qu'entre 65% et 72% de l'état de l'art est collecté avec une telle requête.

Nous décidons d'observer l'impact d'un algorithme prenant en compte directement l'information fournie par le codage IPC. Notre approche est simple, voire simpliste : nous évaluons une stratégie retirant les brevets ne possédant aucun code similaire au brevet utilisé en requête. Dans cette expérience, nous faisons l'hypothèse que le brevet-requête a reçu manuellement des codes IPC. Cette hypothèse est discutable, dans la mesure où l'inventeur déposant un nouveau brevet a rarement procédé à l'attribution des catégories IPC ; toutefois, cette hypothèse est acceptable dans le cas d'une recherche de l'état de l'art faite par un professionnel, employé d'un office national de brevets, dans la mesure où cette démarche est spontanée dans la profession et que de nombreuses formations et quelques outils informatiques sont disponibles pour aider à effectuer de telles tâches (Teodoro et al. 2010). La procédure est évaluée sur le codage partiel (classe et sous-classe), ainsi que sur le codage complet. Dans le Tableau 2, on observe que l'utilisation des descripteurs IPC semble dégrader significativement la capacité du système à générer automatiquement l'état de l'art. Ce résultat est d'autant plus surprenant qu'un gain de précision dépassant 10% (Gobeill et al. 2010) avait été observé avec cette même procédure de filtrage quand une collection de brevets non spécifiquement chimiques est utilisée ! Bien que des études plus fouillées seront probablement nécessaires pour expliciter un tel phénomène, on serait tenté de penser que les méta-données que représentent les classes IPC sont trop génériques pour un domaine scientifique aussi vaste – et donc spécifique – que la chimie. Il conviendrait également d'analyser la distribution des brevets dans la collection que nous utilisons pour ces expériences et celles utilisées par Gobeill et al. 2010, et en particulier les proportions respectives de brevets USPTO vs. EPO dans ces deux collections.

N° DOI : 10.55790/journals/ressi.2010.e1102

Tableau 2 : Précision moyenne (MAP) selon différentes stratégies de filtrage utilisant les codes IPC.

MAP	***IPC filtering* strategy**
Baseline	0.073
4-digits IPC codes	0.074 (-3%)
complete **IPC codes**	0.071 (-8%)

Une hypothèse alternative serait que la qualité et la consistance du codage IPC pour la chimie serait moindre que dans d'autres domaines technologiques. Des expériences sont menées actuellement à la HEG afin de vérifier la qualité relative des modèles de codage IPC dans différents domaines technologiques (Teodoro et al. 2010).

4.3. Réseau bibliographique et expansion automatique

Les derniers contenus informationnels utilisés pour améliorer notre stratégie de génération automatique de bibliographie d'un brevet consistent en substance à privilégier les brevets les plus cités dans l'ensemble de la collection d'une part, et à utiliser des procédures de reconnaissance et d'expansion de termes décrivant des entités chimiques, d'autre part.

Le classement des brevets selon le critère des citations est trivial d'un point de vue computationnel. Il s'agit de classer chaque brevet selon le nombre de citations qu'il obtient dans la collection. Ensuite on combine ce classement statique et indépendant du contenu de la requête avec le classement déjà obtenu à partir du résultat retourné par le moteur de recherche. Cette combinaison linéaire ne doit pas donner trop d'importance aux citations sans quoi chaque requête serait plus ou moins associée au même ensemble de brevets, celui des brevets les plus cités. Empiriquement, nous observons qu'une pondération 10/90 – qui ne donne qu'un dixième d'importance aux citations – fournit les meilleurs résultats. Les performances ainsi obtenues sont impressionnantes avec un gain de + 168%, pour une précision moyenne atteignant 0.179. Ici encore le cas de la chimie est particulier puisque dans une collection de brevets couvrant l'ensemble des technologies, le gain escompté via une telle approche est seulement de +3% (Gobeill et al 2009).

En ce qui concerne l'expansion des noms de composés chimiques, la méthode utilise trois étapes : la reconnaissance (ou étiquetage) des frontières d'entités ; la normalisation de ces entités ; et enfin, l'expansion des entités chimiques reconnues via des synonymes. Cette dernière étape utilise des ressources ontologiques et terminologies gratuites telles que PubChem, DrugBank, OBO, UMLS ou chEBI. Un exemple de la procédure de reconnaissance-normalisation est fourni dans la Figure 2. Cette chaîne de traitement modulaire, appelée ChemTagger, indépendante du moteur lui-même, constitue un puissant outil d'expansion de requête, dont une version simplifiée est disponible publiquement pour démonstration via une interface web (http://eagl.unige.ch/ChemTagger/).

N° DOI : 10.55790/journals/ressi.2010.e1102

Figure 2 : Exemple de procédure de reconnaissance (partie supérieure : texte annoté par des balises XML faisant apparaître chaque composé chimique en rouge) et normalisation (partie inférieure proposant un identifiant unique dans une base de connaissance telle que le MeSH ou PubChem) pour un court passage extrait du brevet # USPTO 6423749. [« A » et « B »]

L'utilisation du ChemTagger lors des évaluations officielles permet une amélioration d'environ 3% pour une précision moyenne de 0.182. Une telle amélioration est certes statistiquement significative mais somme toute assez modeste en regard de la complexité de la procédure d'expansion et de la richesse des ressources terminologiques mobilisées. L'un des problèmes relevé concerne la liste des synonymes fournis par les différentes bases de connaissances utilisées. En effet, afin de ne pas dégrader les performances globales du moteur, seul un nombre restreint de synonymes, si possible les plus communément utilisés, est ajouté automatiquement. Or, pour utiliser pleinement le ChemTagger, il conviendrait plutôt de privilégier un usage interactif, permettant à l'expert de valider et ainsi de contrôler l'expansion. Alternativement, on pourrait essayer de générer un enrichissement automatique de la liste de synonymes avec des informations permettant d'estimer la qualité d'un synonyme particulier, comme proposé récemment par Yepes 2009.

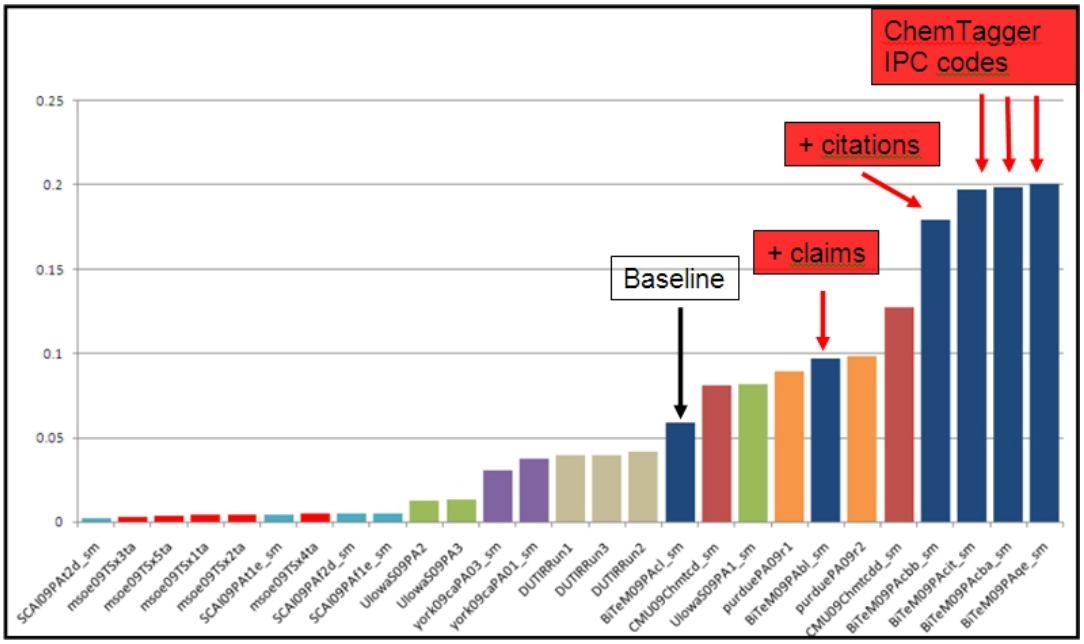
5. Discussion

En termes de comparaison internationale, notre modèle de recherche d'informations obtient des résultats plus que satisfaisants, comme l'illustre la Figure 3, où l'on voit quatre de nos modèles se placer en tête des compétitions officielles TREC 2009. Ces résultats, bien que réjouissants, doivent toutefois être relativisés si l'on considère que nos précédentes participations tendent à montrer que la première année d'une tâche présente souvent les

N° DOI : 10.55790/journals/ressi.2010.e1102

résultats les plus contrastés. Ensuite, les rapports techniques publiés par les participants tendent à faire converger les meilleures approches après quelques itérations autour d'un plafond de précision qui représente l'optimal de la technologie à un moment donné.

Figure 3 : résultats officiels de la compétition TREC 2009 mesuré par la précision des trente premiers brevets retournés (precision@30), extrait de Lupu et. al., 2009).



Notre second constat porte sur l'idée reçue que l'avancée technologique des différents moteurs de RI du web et notamment du plus populaire d'entre eux, va peu à peu remplacer les tâches traditionnellement confiées aux professionnels de la recherche d'informations dans plus ou moins toutes les composantes de leurs activités, dont l'indexation matière, et qu'il est temps de former les futurs professionnel à des tâches alternatives, telles que la médiation (Galaup 2010) ou la redocumentarisation (Salaün 2008) dont la traduction en terme d'activités pratiques nous semble relativement peu claire. Si l'utilité de tâches telles que l'indexation-matière, et la maintenance de ressources terminologiques qu'elle nécessite, est questionnable, il est important que le questionnement ait lieu sur la base d'observations solides. A ce jour, l'importance du rôle des descripteurs assignés manuellement pour effectuer des recherches d'information est variable. Ainsi, dans les bibliothèques de brevets généralistes (Gobeill et al. 2009b) et les bibliothèques digitales médicales (cf. Ruch et al. 2005 ; Abdou et al. 2006), l'importance des descripteurs semble confirmée avec des gains de précision significatifs supérieurs à 5% ($p<0.01$). Inversement, le rôle de tels descripteurs dans le cas des brevets de chimie utilisés dans nos expériences n'est pas confirmé. En d'autres termes, la question n'est pas pour ou contre l'indexation-matière et les thésaurus mais bien de se donner les moyens méthodologiques de mesurer quantitativement l'apport d'un thésaurus particulier, utilisé pour indexer une collection particulière, permettant de répondre à un type de question particulier.

Devant les tentatives plus ou moins abouties de redéfinition des tâches du professionnel de l'information documentaire, il est impératif de discuter – résultats à l'appui – de l'apport informationnel que représentent les descripteurs attribués manuellement par les professionnels de l'information documentaire. Dans tous les cas, l'utilité d'une tâche (e.g. indexation matière) ou d'une ressource (e.g. thésaurus, descripteurs…) doit être évaluée selon

des méthodes scientifiques dans un environnement contrôlé (requêtes, collections…) ; sans quoi on risque le bavardage ou bien le triomphe de la *doxa*.

Par contraste, l'importance des réseaux de citations, tel que proposés originellement par Eugene Garfield (e.g. Garfield 2006), semble particulièrement critique pour les brevets de chimie. On trouve, en effet, dans la collection TREC des brevets partageants des dizaines, voire des centaines de citations, notamment lorsque les inventeurs sont similaires. En terme d'impact sur l'analyse d'antériorité, cette dimension semble donc fondamentale dans le cas de la chimie, sans qu'il soit, à notre connaissance, possible d'expliquer cette spécificité. Au final, on observe que les thésaurus et autres ontologies propres à la chimie, alliées à des outils de traitement automatique de la langue, peuvent constituer un vivier d'améliorations riches mais relativement complexe à exploiter.

Ces résultats sont donc intéressants à bien des égards, notamment scientifiques, toutefois on peut se demander ce que de telles études amènent à la communauté des recherchistes et autres scientifiques de l'information travaillant dans le domaine de la propriété intellectuelle pour les hautes technologies. Cette question aussi légitime que délicate peut trouver bien des réponses selon que l'on considère le court terme ou le long terme. En ce qui concerne le court terme, certains des résultats présentés ici ne font que quantifier des intuitions déjà connues par les professionnels. Ainsi, le fait que le titre du brevet n'apporte que de l'information marginale pour une recherche en antériorité n'est ni nouveau ni surprenant, ce qui l'est c'est le fait que cette intuition est désormais estimée précisément : chercher dans les titres n'améliore la recherche que de 1% ! A moyen terme, les résultats de ces études seront – et sont probablement déjà pour certaines – reprises par les fournisseurs commerciaux de moteurs de recherche pour les brevets. Mais c'est l'impact à long terme qui est le plus fondamental. En effet, si l'on fait fi des considérations commerciales (marketing, force de vente…), et en reconnaissant que la mise à disposition d'un moteur de recherche dans les brevets nécessite la disponibilité d'une collection de brevets de grande qualité, ce qui nécessite souvent de traiter les sources fournies par les offices de brevets, il demeure rassurant de constater qu'une poignée de chercheurs est capable de développer en 2010 ce qui est apparemment un très bon moteur de recherche spécialisés pour la chimie !

6. Conclusion

Nous avons présenté l'état de la recherche dans le domaine des méthodes de recherche d'informations textuelles appliquées aux corpus de brevets et aux brevets de chimie en particulier. Notre rapport expérimental s'appuie sur la campagne d'évaluation TREC 2009 et à la méthodologie Cranfield dont elle hérite. La tâche présentée consistait à générer automatiquement la bibliographie d'un brevet, dont le contenu était utilisé comme requête. Le système présenté utilise une combinaison de méthodes, dont un moteur de recherche vectoriel pondérant différentiellement les champs composant le brevet, un réseau de citation indépendant de la requête, et des procédures de normalisation lexicale utilisant des modules d'extraction de composés chimiques se basant sur des ressources terminologiques publiques telles que PubChem. Le modèle de recherche proposé a obtenu des résultats hautement compétitifs lors des évaluations officielles TREC.

L'autre conclusion rassurante est qu'à l'ère de l' « open source », il est facile de proposer des moteurs de recherches performants et originaux dans des domaines spécialisés – des niches

diront certains – dont le nombre est à l'heure actuel indéfini sinon infini. Cette dernière observation est plus radicale qu'elle ne semble. En effet, elle est de nature à remettre en cause l'idée reçue d'une omnipotence d'un ou deux acteurs du web sur la recherche d'informations en faisant apparaître au grand jour la pluralité des besoins spécifiques de chaque contenu informationnel, voire de chaque corpus. Plus fondamentalement, en affranchissant partiellement les professionnels de leur dépendance vis-à-vis de l'instrument informatique le plus avancé, désormais gratuit et ouvert, il est possible de se détourner de l'arbre cachant la forêt, afin de contempler cette vaste forêt de *Contenu* ! C'est le contenu lui-même, auquel il faut désormais absolument avoir accès, qui conditionne l'accès à l'information. C'est encore ce contenu qui conditionne la valeur ajoutée d'un moteur, dont le développement ne représente qu'un coût marginal(1). On pourrait d'ailleurs suggérer aux grands offices de brevets (EPO, WIPO, USPTO…) la mise en place de services de diffusion des collections de brevets de plus grande qualité. Ainsi la concurrence entre les acteurs commerciaux ne se ferait-elle plus sur leur capacité à uniformiser – parfois manuellement – des sources de faible qualité mais bien à développer de bons moteurs de recherche. Les offices pourrait notamment définir des standards de rédaction plus exigeants, voire fournir des modèles documentaires (MS-Word, Latex), comme le font par exemple les grands éditeurs de journaux.

Enfin, il convient de concéder que notre étude portait exclusivement sur la recherche d'informations textuelles, alors que d'autres modalités pourraient opportunément être étudiées dont la recherche par structure, très utile dans le monde des chimistes. De même, d'autres tâches de recherche d'information sont actuellement à l'étude dans notre laboratoire, dont l'assignement automatiques de codes IPC (Teodoro et al. 2010), la recherche multilingue (Gobeill et al. 2010) ou les tâches de questions-réponses (Pasche et al. 2009). Un tel ensemble de tâches de recherche, apparemment hétérogènes, constituerait pourtant une fois intégrée au sein d'une même interface un outil original et puissant permettant d'interconnecter des fonctionnalités et des sources de connaissances éparses et distribuées : en chimie (e.g. Structures Markush, PubChem), pharmacologie (DrugBank), biologie moléculaire (e.g. Gene Ontologie, Swiss-Prot) ou clinique (e.g. Medical Subject Headings, OMIM).

7. Remerciements

Je tiens à remercier les membres de mon équipe pour l'excellence de leur engagement et la créativité de leurs développements : Arnaud Gaudinat, Julien Gobeill, Emilie Pasche, Douglas Teodoro, Dina Vishnyakova.

8. Disponibilités

Pour les « PatOlympics », organisés ce printemps par l'IRF (Information Retrieval Facility) à Vienne, le groupe a obtenu une première place dans la catégorie recherche dans les brevets de chimie (ChemAthlon) et le prix du jury pour l'interface utilisateur, cf. http://www.ir-facility.org/events/irf-symposium/2010/patolympics pour un compte-rendu et des résultats (http://patolympics.ir-facility.org/PatOlympics/scoreboard.html) officiels. Une version du moteur développé pour la compétition TREC, modifiée pour notre participation aux « PatOlympics » (ainsi appellé TWINC pour « To WIN ChemAthlon »), est disponible en ligne : http://casimir.hesge.ch/ChemAthlon/index.html. Une version simplifiée du ChemTagger et de notre catégoriseur automatique de codes IPC (IPCCat) peuvent être respectivement testés en suivant les liens correspondants sur la page « Resources » du groupe : http://eagl.unige.ch/bitem/.

NOTES

(1) Environ trois chercheurs équivalent plein temps ont travaillé durant six semaines pour développer les stratégies de recherche utilisées pour TREC 2009.

BIBLIOGRAPHIE

Abdou S, Savoy J. (2006): Searching in Medline: Query expansion and manual indexing evaluation. *Information Processing & Management*. Volume 44, Issue 2, March 2008, Pages 781-789

Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS. The Universal Protein Resource (UniProt). *Nucleic Acids Res.* 2005 Jan 1;33.

Brin S, Page L (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. *WWW* 1998.

Capurro, R, Hjørland, B (2003): The Concept of Information. In: Annual Review of Information Science and Technology Ed. B. Cronin, Vol. 37 (2003) Chapter 8, pp. 343-411.

Criscuolo P and Verspagen B, "Does it matter where pa-tent citations come from? Inventor versus examiner cita-tions in European patents", *Research Policy*, vol.37, pp 1892-1908, 2008.

Dolamic D, Savoy J (2010): When stopword lists make the difference. *JASIST* 61(1): 200-203 (2010)

Galaup X, Plaçons la médiation et non les collections au coeur de notre métier in *Archimag*, 1er Mars 2010.

Gobeill J, Teodoro D, Pasche E, Ruch P (2010) Exploring a Wide Range of Simple Pre and Post Processing Strategies for Patent Searching in CLEF IP 2009. *CLEF* 2009

Gobeill J, Teodoro D, Pasche E, Ruch P (2010) Report on the TREC 2009 Experiments: Chemical IR Track. *TREC* 2009.

Garfield E (2006): "The history and meaning of the journal impact factor". *JAMA* 295 (1): 90–3

Grivell, L. "Mining the bibliome: searching for a needle in a haystack?". *EMBO Reports* (3): 200–203.

Kuhn, TS. The Structure of Scientific Revolutions. Chicago: University of Chicago Press, 1962.

Lupu M, Huang J, Piroi F, Zhu J, Tait J. Overview of the TREC 2009 Chemical IR Track. *TREC* 2009.

van Rijsbergen CJ, Robertson SE and Porter MF, 1980. New models in probabilistic information retrieval. London: British Library. (British Library Research and Development Report, no. 5587).

Ruch P (2006): Automatic Assignment of Biomedical Categories: Toward a Generic Approach. *Bioinformatics*, 22(6):658-64, 2006.

Salaün JM : Web, texte, conversation et redocumentarisation, *JADT* 2008.

Sternitzke C, "Reducing uncertainty in the patent applica-tion procedure – insights from malicious prior art in Euro-pean patent applications", *World patent Information*, vol.31, pp 48-53, 2009.

Teodoro D, Pasche E, Vishnyakova D, Gobeill J, Ruch P, Lovis C (2010) Automatic IPC Encoding and Novelty Tracking for Effective Patent Mining. *NTCIR* 2010.

Tseng YH, Wu YJ (2008): A study of search tactics for patentability search: a case study on patent engineers. *PaIR* 2008: 33-36

Yepes A (2009): Ontology Refinement for Improved Information Retrieval in the Biomedical Domain, PhD Thesis, University Jaume I, Castellon, Spain.