F

OMPI

IPC/WG/16/2.
ORIGINAL : anglais
DATE : 20 novembre 2006

ORGANISATION MONDIALE DE LA PROPRIÉTÉ INTELLECTUELLE
GENÈVE

UNION PARTICULIÈRE POUR LA CLASSIFICATION INTERNATIONALE DES BREVETS (UNION DE L'IPC)

GROUPE DE TRAVAIL SUR LA RÉVISION DE LA CIB

Seizième session
Genève, 27 novembre – 1er décembre 2006

THÉMATISATION AUTOMATIQUE DES DOCUMENTS CLASSÉS DANS LES ENDROITS RÉSIDUELS

Document établi par le Secrétariat

1. À sa quinzième session, tenue en juin 2006, le Groupe de travail sur la révision de la CIB a noté que, dans le cadre de la maintenance ordinaire des groupes principaux résiduels, le Bureau international examinerait les possibilités d'utiliser des solutions automatisées pour mettre en évidence les nouvelles technologies dans les groupes principaux résiduels, en regroupant les documents et en identifiant des mots clés permettant de distinguer les différents groupes.

THÉMATISATION AUTOMATIQUE ET ANALYSE DES BESOINS

2. La thématisation automatique consiste à grouper des objets en fonction de leur proximité, définie par une notion spécifique de distance. Ces objets peuvent être des documents ou des concepts contenus dans ces documents. Les groupes ainsi obtenus sont appelés clusters.

3. Pour identifier d'éventuelles entrées nouvelles dans la CIB dans le cadre de la maintenance des groupes principaux résiduels, il convient d'examiner, dans l'espace d'information des sous-classes de la CIB :

– la répartition des concepts dans les documents déjà classés dans un groupe principal résiduel donné;

– la répartition de ces documents; et

– la répartition des différents clusters.

4. L'analyse réalisée dans le cadre du projet CLAIMS a montré que les solutions de thématisation automatique entièrement automatisées ou sans supervision (consistant à grouper des documents sans instructions de l'utilisateur ou sans germe) tendent à grouper les documents en fonction de termes discriminants qui ne sont pas nécessairement intéressants aux fins de la révision de la CIB. L'élaboration de solutions automatiques de ce type constituerait donc un objectif ambitieux en termes d'investissement et de temps. Il est plus approprié et réaliste d'envisager, dans un premier temps, une assistance informatique fondée sur les techniques existantes et leur intégration dans une solution semi-automatisée. Étant donné que les techniques de thématisation automatique purement statistiques risquent de rencontrer certaines limites en cas de proximité linguistique (synonymes), il conviendrait d'envisager également la possibilité d'appliquer des techniques complémentaires fondées sur la proximité sémantique.

5. La solution visée repose sur une assistance informatique permettant de proposer, pour une collection donnée de documents de brevets (déjà classés au niveau des groupes principaux de la CIB), des subdivisions de cette collection (clusters), ainsi que les concepts représentatifs de chacun de ces clusters (centroïde), en tenant compte des éléments suivants :

a) homogénéité des concepts au sein des subdivisions proposées (problème des synonymes et des termes ou expressions sémantiquement proches);

b) proximité des documents ou proximité des concepts : les documents de brevet contiennent généralement différents concepts. Thématisation automatique mixte lorsqu'il s'agit d'agréger non seulement les documents, mais également les concepts;

c) thématisation automatique dans le contexte d'une structure de classement hiérarchique (la CIB) et pas uniquement par partition : thématisation automatique hiérarchique (par agglomération) par opposition à thématisation automatique partitionnelle (par division);

d) la solution devrait être de préférence répétitive, c'est-à-dire produire le même résultat à chaque itération (certaines techniques de thématisation automatique ne sont pas répétitives);

e) la subdivision pourrait être guidée soit par un nombre maximal de clusters, soit par un algorithme de convergence (agrégation jusqu'à ce que l'adjonction d'un nouveau cluster n'ajoute pas d'information suffisante);

f) utilisation éventuellement itérative et combinée d'algorithmes complémentaires et d'interventions humaines lorsque c'est nécessaire. Définition du niveau acceptable d'intervention humaine;

g) utilisation de "concepts répulsifs" au cours de la thématisation automatique : dans le cas des groupes principaux résiduels de la CIB, les documents classés dans ces groupes n'ont, par définition, pas pu être classés dans d'autres groupe principaux de la même sous-classe. Il peut être intéressant d'utiliser cette information, dans le cadre de la solution retenue, afin que les concepts clés proposés pour les subdivisions de ce groupe principal résiduel soient dans la mesure du possible obtenus par déduction des concepts représentatifs figurant dans d'autres groupes principaux de la même sous-classe. Cela permettrait d'éviter que les subdivisions proposées du groupe principal résiduel reprennent certaines parties des groupes principaux existants, voire qu'elles indiquent un classement incorrect des documents dans les groupes principaux résiduels.

SCÉNARIO DE RÉFÉRENCE PROPOSÉ

6. Le scénario ci-après peut être utilisé comme base de discussion et comme référence pour apporter d'éventuelles améliorations. Il consiste en plusieurs itérations des trois étapes ci-après :

Première étape : détermination des clusters possibles en utilisant un mode de thématisation automatique hiérarchique (agglomérative) :

– à partir de l'extraction des termes, regroupements de termes sémantiquement proches (par exemple synonymes), identification de structures linguistiques (centroïdes bruts);

– distribution statistique de concepts et de documents;

– distance entre les concepts, entre les documents et entre les clusters.

Deuxième étape : analyse des éléments communiqués par les experts techniques de la CIB :

– analyser la distance entre les clusters de documents et les clusters de concepts;

– confirmer la terminologie ou proposer des modifications pour la définition des "centroïdes perfectionnés" à utiliser comme germes pour la phase de confirmation;

– en tant qu'option possible, indication du diamètre maximum du cluster pour l'étape suivante.

Troisième étape : essai et confirmation au moyen de la thématisation automatique partitionnelle supervisée (divisive) (par exemple algorithme Quality Threshold Clustering utilisant un diamètre maximal indiqué ci-dessus ou algorithme fuzzy C-means) de façon à avoir des informations sur le degré d'appartenance aux clusters :

– thématisation automatique supervisée utilisant les centroïdes perfectionnés mentionnés ci-dessus comme germes;

– distribution statistique des documents.

ORIENTATION PROPOSÉE POUR LA MISE EN ŒUVRE

7. Étant donné qu'il existe plusieurs techniques complémentaires susceptibles d'être utilisées et qu'une incertitude existe quant aux coûts et avantages associés aux divers modes de mise en œuvre possibles, il est proposé d'adopter une approche en deux étapes.

Phase d'analyse et de prototypage

8. Pendant cette phase, les hypothèses ci-dessus seront analysées de façon plus approfondie, les algorithmes et les scénarios confirmés sur la base du prototypage et des essais réalisés sur des groupes principaux résiduels retenus comme échantillons dans lesquels un nombre suffisant de documents sont classés et sont donc déjà considérés comme susceptibles de faire l'objet d'une révision.

9. La disponibilité de corpus de test pour tester la thématisation automatique pendant cette phase dépendra de la disponibilité de corpus utilisés aux fins de remettre à niveau l'outil d'aide au classement dans la CIB (IPCCAT) en français et en anglais.

10. La préférence devrait être donnée à la mise en œuvre d'une solution basée sur les collections de brevets faisant l'objet d'un classement actualisé. L'identification et la disponibilité d'une collection de ce type (par exemple, extraits de la MCD et de la DOCDB) aux fins de l'analyse et d'une éventuelle application devraient aussi faire l'objet d'une attention particulière.

11. Les incidences financières seront déterminées pour les différentes options résultant de la phase d'analyse.

Mise en œuvre

12. À partir de l'étude ci-dessus, y compris l'évaluation des incidences en termes de ressources, le comité d'experts de la CIB pourra prendre des décisions en ce qui concerne la mise en œuvre et sa dimension définitive.

[Fin du document]