E

WIPO

IPC/WG/16/2
ORIGINAL: English
DATE: November 20, 2006

WORLD INTELLECTUAL PROPERTY ORGANIZATION
GENEVA

SPECIAL UNION FOR THE INTERNATIONAL PATENT CLASSIFICATION (IPC UNION)

IPC REVISION WORKING GROUP

Sixteenth Session
Geneva, November 27 to December 1, 2006

CLUSTERING OF DOCUMENTS CLASSIFIED IN RESIDUAL PLACES

prepared by the Secretariat

1. At its fifteenth session, held in June 2006, the IPC Revision Working Group noted, in the framework of the regular maintenance of residual main groups, that the International Bureau would investigate whether automatic ways of identifying new technologies could be applied in residual main groups, e.g., by clustering documents and identifying keywords that distinguish these clusters.

CLUSTERING AND REQUIREMENTS ANALYSIS

2. Clustering is defined as the action to group objects according to their proximity defined by a specific notion of distance. These objects can be documents and/or concepts inside these documents. These groups are called clusters.

3. For the purpose of identifying possible new IPC entries in the case of IPC residual main groups' maintenance, it is necessary to consider in the IPC subclass information space:

– the distribution of concepts in documents already classified in a particular residual main group;

– the distribution of these documents; and

– the distribution of various clusters.

4. Analysis conducted during the CLAIMS project showed that fully automatic or non-supervised clustering solutions (i.e., clustering documents without user guidance or seed), tend to aggregate documents according to discriminating words which are not necessarily interesting for the purpose of the IPC revision. Developing such automatic solutions is therefore ambitious and would require substantial investments and time. It is more appropriate and realistic to consider, as a first target, computer assistance provided through existing techniques and their integration in a semi-automated solution. As pure statistics-based clustering techniques may show some weaknesses with regard to language proximity (e.g., synonyms), consideration should also be given to complementary semantic proximity techniques.

5. The targeted solution is computer assistance where, for a given collection of patent documents (already classified in the IPC at main group level), subdivisions of this collection (clusters) are proposed, as well as the representative concepts for each cluster (centroid), taking the following into consideration:

(a) homogeneity of concepts within proposed subdivisions (problem of synonyms and semantically close terms or expressions);

(b) document proximity versus concept proximity: patent documents typically include different concepts. Use of co-clustering where not only documents but also concepts are clustered;

(c) clustering in the context of a hierarchical classification structure (the IPC), not only partitioning: hierarchical (agglomerative) versus partitional (dividing) clustering;

(d) the solution should preferably be repetitive, i.e., produce the same result with each run (some clustering techniques are not repetitive);

(e) subdivision could be guided either by a maximum number of clusters or by a converging algorithm (clustering until adding one more cluster does not add sufficient information anymore);

(f) possibly iterative and combined use of complementary algorithms and of human interventions where necessary. Definition of the acceptable level of human interventions;

(g) use of “repulsive concepts” during clustering: in the case of IPC residual main groups, by definition documents classified in residual main groups could not be classified in other main groups of the same subclass. This information may be useful to implement, as an option of the solution used, so that key concepts proposed for subdivisions of this residual main group be as far as possible – repulsed – from representative concepts in other main groups of the same subclass. This would avoid the situation that proposed subdivisions of the residual main group be some of the existing main groups or alternatively may show incorrect classifications of documents in residual main groups.

PROPOSED REFERENCE SCENARIO

6. As a reference, the following scenario can be used for discussion and potential improvements. It consists of several iterations of the three following steps:

Step 1: identification of potential candidate clusters using hierarchical (agglomerative) clustering:

– based on terms extraction, grouping of semantically close terms (e.g., synonyms), identification of linguistic patterns (raw centroids);

– statistical distribution of concepts and of documents;

– distance between concepts, between documents and between clusters.

Step 2: analysis of the submission by IPC technical experts:

– analyze distance between document clusters and concept clusters;

– confirm terminology and/or propose amendments for the definition of “refined centroids” to be used as seeds for the confirmation phase;

– as an option, indication of the maximum diameter of the cluster for the next step.

Step 3: test and confirmation using supervised partitional (dividing) clustering (e.g., Quality Threshold Clustering using a maximum diameter indicated above or Fuzzy C-means) so as to have information about the degree of belonging to clusters:

– supervised clustering using the above-mentioned refined centroids as seeds;

– statistical distribution of documents.

PROPOSED IMPLEMENTATION APPROACH

7. As there are a number of complementary techniques that can be used and uncertainty on cost/benefits of the various possible options for implementation, it is proposed to adopt a staged approach in two steps.

Analysis and Prototyping Phase

8. During this phase the above assumptions will be analyzed in more detail, algorithms and scenario confirmed on the basis of prototyping and tests conducted on sample residual main groups where a large enough number of documents are classified and thus already identified as candidates for revision.

9. Availability of test corpora to test clustering during this phase will depend on the availability of corpora used for the purpose of retraining the IPCCAT categorization assistance tool in English and French.

10. Preference should be given to implementing a solution on the basis of patent collections with up-to-date classification. Identification and availability of such collection (e.g., MCD and DOCDB extracts) for the purpose of the analysis and possible implementation should also be carefully considered.

11. Financial implications will be associated with different options resulting from the analysis phase.

Implementation

12. On the basis of the above study, including assessment of resource implications, the IPC Committee of Experts can make decisions on implementation and its final scope.

[End of document]