Internal distribution code:

(A) [-] Publication in OJ
(B) [-] To Chairmen and Members
(C) [-] To Chairmen
(D) [X] No distribution

Datasheet for the decision of 11 April 2014

Case Number: T 0233/09 - 3.5.07

Application Number: 04818002.0

Publication Number: 1709557

IPC: G06F17/30

Language of the proceedings: EN

Title of invention:
System and method for comparative analysis of textual documents

Applicant:
Hewlett-Packard Development Company, L.P.

Headword:
Comparative analysis of text documents/HEWLETT-PACKARD DEVELOPMENT

Relevant legal provisions:
EPC Art. 52(2), 52(3), 56

Keyword:
Exclusion from patentability - main request (yes) - first auxiliary request (yes) - second auxiliary request (yes) Inventive step - third auxiliary request (no)

Decisions cited:
T 0388/04, T 1316/09

Catchword:

Europäisches Patentamt
European Patent Office
Office européen des brevets

Beschwerdekammern
Boards of Appeal
Chambres de recours

European Patent Office
D-80298 MUNICH
GERMANY
Tel. +49 (0) 89 2399-0
Fax +49 (0) 89 2399-4465

Case Number: T 0233/09 - 3.5.07

D E C I S I O N
of Technical Board of Appeal 3.5.07
of 11 April 2014

Appellant: (Applicant)	Hewlett-Packard Development Company, L.P. 11445 Compaq Center Drive West Houston, TX 77070 (US)
Representative:	Lawman, Matthew John Mitchell EIP Fairfax House 15 Fulwood Place London, WC1V 6HU (GB)
Decision under appeal:	**Decision of the Examining Division of the European Patent Office posted on 27 June 2008 refusing European patent application No. 04818002.0 pursuant to Article 97(2) EPC.**

Composition of the Board:

Chairman: R. Moufang
Members: R. de Man
P. San-Bento Furtado

Summary of Facts and Submissions

I. The applicant (appellant), which at the time was Electronic Data Systems Corporation, filed an appeal against the decision of the Examining Division refusing European patent application No. 04818002.0.

II. With effect from 18 May 2009 the application was transferred to Hewlett-Packard Development Company, L.P., which thereby obtained the status of appellant.

III. In the contested decision, reference was made to the documents

D1: Besançon R., Rajman M., Chappelier J.-C.: "Textual Similarities based on a Distributional Approach" (1999); and
D2: Hotho A., Staab S., Stumme G.: "Ontologies Improve Text Document Clustering" (2003).

The Examining Division came to the conclusion that the main request did not comply with Article 123(2) EPC and that the subject-matter of claims 1, 12 and 14 of the auxiliary request was new neither over document D1 nor over document D2.

IV. With the statement of grounds of appeal, the appellant filed a main request and two auxiliary requests. The appellant requested oral proceedings in case the Board deemed the main request unallowable.

V. The Board appointed oral proceedings. In a communication accompanying the summons, it provided its provisional opinion. The subject-matter of claim 1 of all requests appeared to be excluded from patentability within the meaning of Article 52(2) and (3) EPC.

Interpreting these claims as defining computer-implemented methods, their subject-matter appeared to lack an inventive step in view of the cited prior art, but also without reference to a document. In addition, claim 1 of the first auxiliary request appeared to lack novelty.

VI. With a letter dated 10 March 2014, the appellant filed a third auxiliary request and confirmed that it maintained each of the main, first and second auxiliary requests.

VII. With a letter dated 31 March 2014, the appellant informed the Board that it would not attend the oral proceedings.

VIII. Oral proceedings were held on 11 April 2014 in the absence of the appellant. At the end of the oral proceedings, the chairman announced the decision of the Board.

IX. The appellant requested that the decision under appeal be set aside and that a patent be granted on the basis of the main request, or in the alternative, of one of the first to third auxiliary requests.

X. Claim 1 of the main request reads as follows:

"A method of comparing the semantic content of two or more documents (710, 801) comprising:

accessing two or more documents (710, 801);

performing a linguistic analysis on each document;

outputting a quantified representation of the semantic content of each document; and

comparing the quantified representations using a defined metric, wherein the quantified representation

of a semantic content is a semantic vector that has multiple components (510, 520, 530, 540), characterized in that each component of the semantic vector has at least:

a word or phrase appearing in the document or a synonym of the word or phrase;

a weighting factor related to an importance of the word or phrase or synonym in the document; and

a frequency value."

XI. Claim 1 of the first auxiliary request reads as follows:

"A method of comparing the semantic content of two or more documents (710; 801), comprising:

accessing the two or more documents (710; 801);

performing a linguistic analysis on each document (710; 801); and

outputting a quantified representation of the semantic content of each document, wherein the quantified representation is a semantic vector having multiple components (510, 520, 530, 540); and

comparing the quantified representations using a defined metric characterised in that the metric generates a measure of semantic distance between the two or more documents."

XII. Claim 1 of the second auxiliary request reads as follows:

"A method of comparing the semantic content of two or more documents (710; 801), comprising:

accessing the two or more documents (710; 801);

performing a linguistic analysis on each document (710; 801); and

outputting a quantified representation of the semantic content of each document;

wherein the quantified representation is a semantic vector having multiple components (510, 520, 530, 540) comprising weighting factors associated with a word or phrase or a synonym of said word or phrase and/or frequency values;

comparing the quantified representations using a defined metric characterised in that the metric generates a measure of semantic distance between each vectors comprising calculating a difference in weighting factors or frequencies."

XIII. Claim 1 of the third auxiliary request reads as follows:

"A computer-implemented method of comparing the semantic content of two or more documents (710, 801) comprising:

accessing text data representative of two or more documents (710, 801);

performing text analysis on text data for each document, including:

tagging domain words or phrases in said text data, domain words or phrases being identified based on a list of domain words or phrases, the list of domain words or phrases being derived from an analysis of commonly occurring words or phrases in a set of domain documents that exclude words in a word exclusion list;

tagging redundant words in said text data, redundant words being identified based on words in the word exclusion list;

performing a frequency analysis on occurrences of each domain word or phrase that are not tagged as redundant;

assigning a weighting factor to occurrences of each word or phrase that are not tagged as redundant, the weighting factor being set based on one or more word relevance rules, the word relevance rules comprising rules that apply a weight based on one or more of domain word properties and word order properties;

outputting a semantic vector representative of the text data for each document that has multiple components (510, 520, 530, 540), including at least:

a word or phrase appearing in the text data or a synonym of the word or phrase;

a weighting factor associated with the word or phrase or synonym in the text data; and

a frequency value associated with the word or phrase or synonym in the text data; and

outputting a defined distance metric representative of the similarity of the two or more documents, the defined distance metric being calculated based on the set of outputted semantic vectors."

Reasons for the Decision

1. The appeal complies with the provisions referred to in Rule 101 EPC and is therefore admissible.

2. *Overview*

The application is concerned with the comparative analysis of textual documents by creating and comparing semantic vectors.

A semantic vector is a mathematical object intended to represent the semantic content of a document. It consists, for example, of a number of words occurring

in the document together with their frequencies of occurrence. Instead of frequencies, or in addition to frequencies, a semantic vector may include for each word a weight that somehow relates to the importance or relevance of that word in the document. The application gives several examples of "Word Relevance Rules" for determining the weight of a word.

Semantic vectors are compared by applying a "semantic distance metric", which is a mathematical formula that, based on the frequencies and/or weights contained in two semantic vectors, calculates a number intended to be representative of the "semantic closeness" of the two documents to which the semantic vectors correspond.

The application discloses various applications of the comparison method, one example being the search for prior-art documents which are semantically close enough to a set of patents to potentially invalidate one or more of them.

3. *Main request, first auxiliary request, second auxiliary request - Article 52(2) and (3) EPC*

3.1 Independent claim 1 of the main request defines a "method of comparing the semantic content of two or more documents". It recites steps of accessing the documents, performing a linguistic analysis on each document, outputting a quantified representation of the semantic content of each document, and comparing the quantified representations using a defined metric. None of these steps inherently implies the use of technical means.

In particular, the step of "outputting" in the context of claim 1 merely defines that a quantified

representation of the semantic content of each document is produced as input for the step of "comparing". This quantified representation takes the form of a semantic vector, which is a mathematical object which in principle does not require a physical manifestation.

Claim 1 hence consists of a series of abstract activities, relating to a mixture of linguistics and mathematics, and which in principle may be carried out mentally. Although the steps of claim 1 certainly *may* be performed using technical means, the mere possibility of making use of unspecified technical means for performing an activity is not sufficient to lend that activity technical character (see decision T 388/04 - Undeliverable mail/PITNEY BOWES, OJ EPO 2007, 16).

The Board therefore concludes that the subject-matter of claim 1 relates to a mental method as such and is hence excluded from patentability under Article 52(2) and (3) EPC.

3.2 The same observations apply to independent claim 1 of the first auxiliary request and to independent claim 1 of the second auxiliary request.

The subject-matter of claim 1 of both the first and the second auxiliary request is therefore excluded from patentability as well (Article 52(2) and (3) EPC).

4. The objection under Article 52(2) and (3) EPC in respect of claim 1 of the main request and the first and second auxiliary requests was communicated to the appellant in the communication accompanying the summons. The appellant chose to address it only through the filing of a third auxiliary request. Independent

claim 1 of this request defines a computer-implemented method, and hence does avoid exclusion.

5. For completeness, the Board observes that in the communication accompanying the summons an inventive step objection was raised in respect of claim 1 of the main request and the first and second auxiliary requests when interpreted as defining computer-implemented methods. In addition, a novelty objection was raised in respect of claim 1 of the first auxiliary request. The appellant has not addressed these objections, and the Board still considers them justified.

6. *Third auxiliary request - Article 56 EPC*

6.1 As a preliminary remark, the Board notes that it may be questioned whether any of the features of the computer-implemented method of claim 1 contributes to a technical effect going beyond the mere automation of a non-technical task. Indeed, the method of claim 1 is defined in functional terms, i.e. not in terms of a concrete technical implementation, and any overall effect appears to lie in the non-technical field of linguistics. See also decision T 1316/09 of 18 December 2012, in which a claim involving the calculation of a text-mining score as a measure of the similarity between two text documents was not considered to produce any relevant technical effect. However, the Board chooses to proceed starting from document D1.

6.2 Document D1 relates to computational techniques for determining textual similarity of documents, see abstract and section 1, second paragraph. The Board

therefore considers document D1 to be a suitable starting point for the assessment of inventive step.

6.3 Document D1 is specifically concerned with the identification of relevant documents in a database on the basis of a textual query. Section 1, second paragraph, explains that this problem can be viewed as the search for the documents the most similar to the query, which search can be carried out through the computation of textual similarities between the query and each of the documents in the database.

Section 2.1.1 discusses the "Vector Space model", which represents each document as a vector in which weights are assigned to terms occurring in the document. The terms used to index the documents are chosen to be as discriminative as possible. The weight of a term may simply be the number of occurrences of the term in the document, referred to as *occurrence frequency*, or it may take into account the term's importance within the entire document collection. More weight may for example be given to terms that rarely occur within the collection.

Section 2.1.2 discloses that the similarity between a document and a query may be measured by calculating the cosine of the angle between their two vectors or by other similarity measures such as the χ^2 distance.

6.4 Claim 1 of auxiliary request 3 is directed to a computer-implemented method of comparing the semantic content of two or more documents, essentially by determining textual similarity. It first performs text analysis to produce for each document a representative semantic vector. It then determines the similarity of

the two or more documents by calculating a "defined distance metric" based on the semantic vectors.

6.5 According to the description on page 6, lines 1-3, the term "document" is to be understood in its most expansive sense and includes any quantity of text in any format that can be subjected to linguistic analysis. The textual queries of document D1 are hence documents within the meaning of claim 1.

6.6 Furthermore, given that the context of the computation of textual similarities in document D1 is the retrieval of documents from large textual databases, it is evident that document D1 contemplates a computer-implemented method of comparing the semantic content of two or more documents, wherein the documents are provided in the form of representative (digital) text data.

6.7 Document D1 further determines a measure of similarity between documents by calculating a value based on the semantic vectors of those documents, for example by applying the χ^2 distance metric.

In the statement of grounds of appeal, the appellant submitted that document D1 does not disclose a measure of semantic distance between vectors, "i.e. involving subtracting one vector from another vector to determine the distance between the vectors". However, the Board sees no reason why the calculation of a distance between two vectors necessarily involves subtracting one vector from the other. The Board therefore considers that the χ^2 distance referred to in document D1 is a measure of distance within the meaning of the claim.

6.8 Document D1 hence discloses a method comprising the following features:

- accessing text data representative of two or more documents;
- performing text analysis on text data for each document;
- outputting a semantic vector representative of the text data for each document that has multiple components, including at least:
  - a word or phrase appearing in the text data;
  - a frequency value associated with the word or phrase in the text data; and
- outputting a defined distance metric representative of the similarity of the two or more documents, the defined distance metric being calculated based on the set of outputted semantic vectors.

6.9 The definition of the semantic vector of claim 1 refers to "a word or phrase appearing in the text data or a synonym of the word or phrase". However, the text analysis steps of claim 1 do not identify any synonyms and do not determine frequency values and/or weighting factors on the basis of the presence of synonyms. The Board therefore considers the expression "or a synonym of the word or phrase" to be an (unclear) optional feature which can hence be ignored in the assessment of inventive step.

6.10 Claim 1 therefore differs from the method of document D1 in that

(a) the semantic vectors include both weighting factors and frequency values

and in features relating to the assignment of weighting factors and frequency values to a selected subset of the terms of a document:

(b) a list of domain words or phrases and a word exclusion list are provided, the list of domain words or phrases "being derived from an analysis of commonly occurring words or phrases in a set of domain documents that exclude words in a word exclusion list";
(c) domain words or phrases in the document that are present in the list of domain words or phrases are tagged;
(d) words in the document that are present in the word exclusion list are tagged as redundant;
(e) a frequency analysis is performed on occurrences of each domain word or phrase not tagged as redundant; and
(f) a weighting factor is assigned to each word or phrase not tagged as redundant, the weighting factor being set based on one or more word relevance rules, the word relevance rules comprising rules that apply a weight based on one or more of domain word properties and word order properties.

The Board notes that, since feature (b) specifies that the "list of domain words or phrases" excludes the words in the "word exclusion list", feature (e) can be read as "a frequency analysis is performed on occurrences of each domain word or phrase".

6.11 *Feature (a)*

6.11.1 Document D1 discloses the assignment of a weighting factor to a term in a document, an example of a

weighting factor being the frequency of the term, i.e. the number of its occurrences in the document. Document D1 does not disclose assigning both a weighting factor and a frequency value.

6.11.2 In the statement of grounds of appeal, the appellant submitted that document D1 did not disclose feature (a), but did not provide reasons as to why this distinguishing feature supported an inventive step.

It could be argued that the presence of both weighting factors and frequency values in semantic vectors leads to a "better" measure of textual similarity. However, claim 1 does not specify a concrete distance metric in which the use of both weighting factors and frequency values is expressed. On the contrary, it follows from the example metrics given in dependent claim 10 that the distance metric of claim 1 covers metrics based solely on frequency values and metrics based solely on weighting factors. Feature (a) therefore does not in fact lead to a "better" measure of textual similarity.

6.11.3 Since claim 1 covers embodiments in which the weighting factors are not used in determining textual similarity of documents, the Board is of the view that their presence in the semantic vectors in addition to the frequency values, as expressed by feature (a), cannot contribute to an inventive step.

Moreover, even if claim 1 had to be understood, in contradiction to dependent claim 10, as covering only distance metrics that take into account both weighting factors and frequency values, the Board considers that, in view of the fact that basing a distance metric on either weighting factors or frequency values is known from document D1, it is obvious to base a distance

metric on values of both types, and, as a consequence, to include values of both types in the semantic vectors.

6.12 *Features (b)-(f)*

6.12.1 Document D1, section 2.1.1, discloses that the terms to be included in the semantic vectors are to be chosen so as to be as discriminative as possible, one possibility being to select terms based on the number of documents in which they occur.

6.12.2 Features (b), (c) and (e) essentially express that the terms to which frequency values are assigned are selected as a set of "commonly occurring words or phrases" in a set of domain documents. Since the restriction to domain documents merely concerns the cognitive content of the documents being analysed, this distinction is non-technical and therefore cannot contribute to an inventive step. In any event, it is obvious to apply the techniques of document D1 to a textual database containing documents in a particular domain.

6.12.3 Features (b), (d) and (f) essentially express that the terms to which weighting factors are assigned are the terms not present in a "word exclusion list". According to document D1, section 2.1.1, terms that are used in many documents are more general and less useful for discrimination than ones that appear in very few documents. The Board therefore considers it an obvious possibility to exclude such words from consideration by including them in a "word exclusion list". Similarly, it is obvious to ensure that such words are not included in the "list of domain words or phrases".

6.12.4 Feature (f) further defines that weighting factors are assigned based on "word relevance rules", including rules that apply a weight based on a "domain word property". Document D1, section 2.1.1, suggests giving more weight to terms that rarely occur within the document collection. Where the document collection consists of documents in a particular domain, the Board considers this a "domain word property".

6.13 From the above considerations it follows that feature (a) relates to an obvious variation of the method of document D1 and that features (b)-(f) define obvious details of this variation. The subject-matter of claim 1 of the third auxiliary request therefore lacks an inventive step within the meaning of Article 52(1) and 56 EPC.

7. Since none of the requests on file is allowable, the appeal has to be dismissed.

Order

For these reasons it is decided that:

The appeal is dismissed.

The Registrar: The Chairman:

I. Aperribay R. Moufang

Decision electronically authenticated