(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
17 March 2011 (17.03.2011)

PCT

(10) International Publication Number
WO 2011/029474 A1

(51) International Patent Classification:
G06F 17/22 (2006.01)

(21) International Application Number:
PCT/EP2009/061713

(22) International Filing Date:
9 September 2009 (09.09.2009)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant *(for all designated States except US)*: **UNIVERSITÄT BREMEN** [DE/DE]; Bibliothekstr. 1, 28359 Bremen (DE).

(72) Inventor; and

(75) Inventor/Applicant *(for US only)*: **MÖHRLE, Martin, G.** [DE/DE]; Waetjenstr. 12, 28213 Bremen (DE).

(74) Agent: **RUMMLER, Felix**; Martiusstr. 5, 80802 Munich (DE).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

[Continued on next page]

(54) Title: DOCUMENT COMPARISON

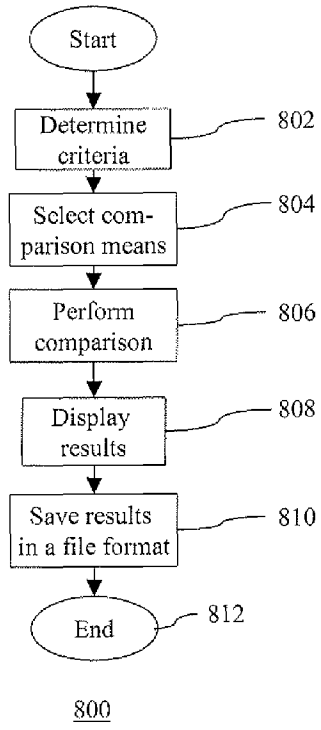


Figure 8

(57) Abstract: A method of comparing first and second documents, the method comprising determining criteria of the comparison, selecting comparison means based on the criteria from a plurality of comparison means, and performing the comparison of the first and second documents by using the selected comparison means. Also disclosed is a method of comparing first and second documents, the first document including or associated with one or more first concepts, the second document including or associated with one or more second concepts, the method comprising displaying a diagram having first and second axes, the first axis corresponding to positions of the first concepts within the first document, the second axis corresponding to positions of the second concepts within the second document, the method further comprising displaying or highlighting one or more points on the diagram, each point at a position on the first axis corresponding to a first one of the concepts in the first document and on the second axis corresponding to a second one of the concepts in the second document, whereby the first concept is identical or similar to the second concept.

WO 2011/029474 A1

Published:

— *with international search report (Art. 21(3))*

DOCUMENT COMPARISON

FIELD OF EMBODIMENTS OF THE INVENTION

Embodiments of the invention relate to document comparison, for example for comparing the text content of documents.

BACKGROUND TO EMBODIMENTS OF THE INVENTION

Document comparison using electronic means is useful, particularly when there are a large number of documents to compare especially when the documents have a similar structure and are segmentable within this structure. Electronic document comparison may use one or more data processing systems to compare text from the documents, the text also being in or being obtainable in electronic form.

For example, an individual may wish to compare two or more patent documents (that is, granted patents, patent applications, provisional applications, utility model applications and the like). There exists an enormous number of published patent documents, making the task of manually comparing the documents, or selecting documents for comparison, complicated. Electronic methods of comparing patent documents or selecting documents for comparison may therefore be useful.

SUMMARY OF EMBODIMENTS OF THE INVENTION

Aspects of embodiments of the invention are set out in the claims.

BRIEF DESCRIPTION OF THE DRAWINGS

Embodiments of the invention will now be described by way of example only with reference to the following figures, in which:

Figure 1 shows an example of determining variables in a complete linkage method according to embodiments of the invention;

Figure 2 shows an example of determining variables in a complete linkage method according to embodiments of the invention;

Figure 3 shows an example of determining variables in a complete linkage method according to embodiments of the invention;

Figure 4 shows an example of determining variables in a reduced linkage method according to embodiments of the invention;

Figure 5 shows an example of determining variables in a wedding linkage method according to embodiments of the invention;

Figure 6 shows an example of determining variables in an integer linkage method according to embodiments of the invention;

Figure 7 shows an example of determining variables in a bounded integer linkage method according to embodiments of the invention;

Figure 8 shows an example of a method for comparing documents according to embodiments of the invention;

Figure 9 shows an example of a diagram for comparing documents according to embodiments of the invention; and

Figure 10 shows an example of a data processing system suitable for use with embodiments of the invention.

DETAILED DESCRIPTION OF EMBODIMENTS OF THE INVENTION

Embodiments of the invention compare two or more documents based on the concepts that are found within the documents. Concepts may comprise, for example, general notions, ideas or subjects found or referred to in the documents. The concepts may then be used to determine one or more numerical values reflecting the similarity of the documents. Additionally or alternatively, the concepts and their locations within the

documents may be used to produce a diagram of concepts that are common to multiple documents, indicating in which part of the documents there are accumulations of similar concepts, thus leading to an in-depth analysis of document relationships.

According to first embodiments of the invention, criteria of a document comparison are determined, and comparison means for comparing the documents are selected based on the criteria. The comparison means may then be used to compare the documents, for example to provide one or more numerical values reflecting the similarity or some of its facets of two or more documents.

A numerical value reflecting the similarity of two documents may use variables based on the concepts within the two documents. A document may include or be associated with one or more concepts. There may be multiple concepts that are identical or similar (for example, use alternative words for the same or similar meaning). Concepts within a document may be predefined or may be extracted from the document. There are various ways of extracting concepts from a document and an example is provided below. Concepts may be determined by extraction from a document or by accessing the predefined concepts, or by other means. Where a concept is referred to as occurring multiple times, having duplicates or being identical to another concept, this should be understood to mean that the multiple concepts are identical or similar to each other.

In a first method of extracting concepts from a document, free morphemes or root forms of words are considered. For example, where a document contains any of the words “machine”, “machines” or “machinery”, these may all be considered to be the single word “machine”. Thus, each word in the document is replaced by its free morpheme. Next, certain words are disregarded. For example, certain words may be likely to appear in many if not most or all of the documents under consideration. In patent documents, for example, the words “figure”, “show”, “embodiment”, “claim” and others may be expected to be in a large proportion of the documents under consideration. These words may be disregarded. The remaining words can be considered to be a list of the concepts in the document.

Once concepts have been determined, variables may be calculated based on the concepts. For example, according to certain embodiments of the invention, up to five variables may be defined. c_i is the number of concepts in a first document (document i). c_j is the number of concepts in a second document (document j). Where there are multiple identical or similar concepts in a single document, these are counted each time they appear, and so the variables c_i and c_j may be higher than the number of unique concepts in the respective documents. The variable $c_{i(j)}$ is the number of concepts in document i that have identical or similar equivalents in document j. $c_{j(i)}$ is the number of concepts in document j that have identical or similar equivalents in document i. c_{ij} is the number of concepts that can be found in both documents. c_{ij} may differ from $c_{i(j)}$ and $c_{j(i)}$, which may depend on a method selected for measuring these variables.

A method is selected for measuring the variables from a number of methods. The selected method may give different results for the variables $c_{i(j)}$, $c_{j(i)}$ and c_{ij} than other methods. The methods differ in the way they consider multiple occurrences of identical or similar concepts in a single document. One method, "complete linkage", is shown in figure 1, which shows an example of concepts in or associated with two example documents. In the complete linkage method, each concept is treated as if it is unique and is counted separately from other concepts, even other identical or similar concepts. As shown in figure 1, document i includes five concepts, A, B, C, E and F. In document i, concept A appears twice, concept B appears three times, concept C appears once, concept E appears once and concept F appears twice. In document j, concept A appears once, concept B appears twice, concept D appears once, concept F appears twice and concept G appears once.

The variables c_i and c_j are counts of the number of concepts in document i and document j respectively, and are 9 and 7 respectively. To calculate the variable c_{ij}, the total number of common concepts is determined. Therefore, for complete linkage, multiple concepts in one document that have identical concepts in the other document are considered multiple times. This is illustrated by lines drawn between identical or similar concepts in figure 1. Thus, the two occurrences of concept A in document i and the single occurrence in document j contribute 2 to the variable c_{ij}, and the three occurrences of concept B in document i and the two occurrences in document j

contribute 6 to c_{ij}, because each occurrence of concept B in document i is considered against each occurrence in document j. Thus, the final value for c_{ij} is 12.

Figure 2 illustrates an example of calculation of the variable $c_{i(j)}$, which indicates the number of concepts in document i that can also be found in document j. Again, multiple occurrences of a concept in document i are considered multiple times, although multiple occurrences of a concept in document j do not affect the value of $c_{i(j)}$. The value for $c_{i(j)}$ in the example shown is 7. Similarly, figure 3 shows an example of calculation of the variable $c_{j(i)}$, which is 5 in the example shown.

A second method for measuring the variables is shown in figure 4. This method, called "reduced linkage", considers multiple occurrences of identical or similar concepts in a document as just one occurrence. Therefore, as shown in figure 4, the variable c_{ij} is determined to be 3. Similarly, the variables $c_{i(j)}$ and $c_{j(i)}$ are also measured to be 3.

A third method for measuring the variables, called "wedding linkage", is shown in figure 5. In this method, for each concept in one document, a match is searched for in the other document. Where a match is found, this contributes to the variable c_{ij} for example, and the matched concepts in both documents can no longer be used. Therefore, for multiple occurrences of a concept to be counted multiple times, there must be multiple occurrences in both documents. For example, as shown in figure 5, there are multiple occurrences of the concept A in document i, but only one in document j, so concept A contributes only once to c_{ij}. On the other hand, there are three occurrences of concept B in document i and two in document j, so the concept B contributes twice to c_{ij}. In the example shown, using the wedding linkage method provides $c_{ij} = 5$, $c_{i(j)} = 5$ and $c_{j(i)} = 5$.

A fourth method, "integer linkage", is shown in figure 6. In this method, multiple concepts are treated as a single concept for calculating the variables as in reduced linkage. However, the number of multiple occurrences is used to provide a weighting to the contribution to the variables of the reduced concepts. In the example shown in figure 6, the weighting given is the number of occurrences of a concept in document i multiplied by the number of occurrences of this concept in document j. For example,

the contribution to c_{ij} by the concept B is 3x2 = 6. This weighting gives results for the variables as $c_{ij} = c_{i(j)} = c_{j(i)} = 12$. However, in alternative embodiments other weighting methods can be used.

A fifth method, "bounded integer linkage", is shown in figure 7. This method is similar to integer linkage as described above. However, in bounded integer linkage, the weighting given to multiple occurrences of a concept in one document is no more than a predetermined maximum number. In the example shown, this maximum is 2, so the weighting given to the three occurrences of concept B in document i does not exceed 2. In the example shown, the contribution to the variables such as c_{ij} by common concepts between the documents is equal to the weighting given to the number of occurrences of the concept in document i multiplied by the weighting given in document j. For example, the contribution by concept B is 2x2 = 4. According to this example, the variables are calculated to be $c_{ij} = c_{i(j)} = c_{j(i)} = 10$, although as for the integer linkage method other ways of calculating the contribution and/or other maximum values can be used.

Once a method of calculating the variables has been chosen and the variables have been calculated, a method is chosen for determining a value that reflects the similarity between the two documents being compared. In certain embodiments of the invention, this comprises choosing one of a number of similarity coefficient formulas. Examples of similarity coefficient formulas are given below. The formulas can be split into two categories: those that give two-sided overlap coefficients, and those that give double single-sided overlap coefficients.

The two-sided overlap coefficient formulas use the variables c_i, c_j and c_{ij}. A number of examples of such formulas are given in table 1 below:

Similarity coefficient	Definition
Jaccard	$\frac{c_{ij}}{c_i + c_j - c_{ij}}$
Sorensen	$\frac{2c_{ij}}{c_i + c_j}$

Sokal & Sneath 2	$\frac{c_{ij}}{2(c_i + c_j) - 3c_{ij}}$
Kulczynski 1	$\frac{c_{ij}}{c_i + c_j - 2c_{ij}}$
Kulczynski 2	$\frac{\frac{c_{ij}}{c_i} + \frac{c_{ij}}{c_j}}{2}$
Cosine	$\sqrt{\frac{c_{ij}}{c_i} \cdot \frac{c_{ij}}{c_j}} = \frac{c_{ij}}{\sqrt{c_i \cdot c_j}}$
Inclusion	$\min\left(\frac{c_{ij}}{c_i}, \frac{c_{ij}}{c_j}\right)$

Table 1: two-sided overlap similarity coefficient formulas

Double single-sided (DSS) overlap coefficient formulas use the variables c_I, c_j, $c_{i(j)}$ and $c_{j(i)}$, and a number of examples are given in table 2 below:

Similarity coefficient	Definition
DSS-Jaccard	$\frac{c_{i(j)} + c_{j(i)}}{c_i + c_j}$
DSS-Inclusion	$\max\left(\frac{c_{i(j)}}{c_i}, \frac{c_{j(i)}}{c_j}\right)$
DSS-Inclusion (extreme variant)	$\frac{\max(c_{i(j)}, c_{j(i)})}{\min(c_i, c_j)}$
DSS-Gamma-Inclusion	$-1 + \left[\left(\frac{c_{i(j)}}{c_i}\right)^{\gamma} + \left(\frac{c_{j(i)}}{c_j}\right)^{1-\gamma}\right], \quad 0 \le \gamma \le 1$

Table 2: double single-sided overlap similarity coefficient formulas

The DSS-Gamma-Inclusion formula includes a weighting variable γ. This variable can be used to balance between two simple one-sided coefficients. For example, to

balance the formula equally between the two one-sided coefficients, γ is chosen to be 0.5.

Of the two-sided overlap similarity coefficient formulas listed above, the Jaccard, Cosine and Inclusion coefficients will be considered further. However, in alternative embodiments, other formulas for the two-sided and DSS overlap similarity coefficients may be used that may or may not be those listed above.

Table 3 below gives results for selected ones of the two-sided overlap similarity coefficient formulas using various methods for determining the variable c_{ij} as identified above. The results provide values reflecting the similarity of the two documents being compared.

Similarity coefficient	**Complete linkage**	**Reduced linkage**	**Wedding linkage**	**Bounded integer linkage**
Jaccard	3	0.43	0.45	1.67
Inclusion	1.7	1	0.71	1.43
Cosine	1.52	1	0.63	1.26

Table 3: two-sided overlap similarity coefficient results

Table 4 below gives results for the double single-sided (DSS) overlap similarity coefficient formulas. For DSS-Gamma-Inclusion, $\gamma = 0.5$.

Similarity coefficient	**Complete linkage**	**Reduced linkage**
DSS-Jaccard	0.75	0.6
DSS-Inclusion	0.78	0.6
DSS-Gamma-Inclusion	0.73	0.55

Table 4: DSS overlap similarity coefficient results

In certain embodiments of the invention, document comparison means comprise or include the formula and the method for determining the variables used by the formula.

The document comparison means are selected based on one or more criteria of the comparison. For example, the criteria of the comparison may include a purpose of the comparison of the documents, an importance of considering duplicate concepts in the documents, a distribution of duplicate concepts and a size distribution of documents. These examples are explained in more detail below.

One of the criteria used in the selection of the document comparison means may comprise a purpose of the document comparison. For example, where one or both of the documents is a patent document, the purpose may comprise a prior art analysis, infringement analysis or patent document similarity mapping. For prior art or infringement analysis, a document of interest may be compared with a plurality of other patent documents. It may be undesirable to miss any patent document that is potentially similar to the document of interest. Therefore, for example, a threshold of 0.2 may be set, and a document that has a similarity coefficient of greater than the threshold may be marked for manual comparison with the document of interest. In this case, selection of the inclusion or DSS-inclusion similarity coefficient formula may be desirable, as the values from these formulas tend to be greater than for other formulas as shown above. Thus, more documents are above the threshold and more documents are marked for manual comparison, reducing the risk that an important document is not marked for manual comparison.

For patent mapping, a m*m matrix of similarity coefficients may be obtained where m is the number of documents being compared. As a result of the large number of coefficients, it may be appropriate to use a more conservative similarity coefficient formula, such as Jaccard or DSS-Jaccard.

The criteria may also include an importance of considering duplicate concepts in the documents. In some cases, for example, there may be rare or unusual concepts in one of the documents being compared, and so selection of document comparison means that puts greater emphasis on multiple occurrences of identical or similar concepts may be desired. Thus, use of the complete linkage method for variable measurement may be appropriate, and/or use of a two-sided overlap similarity coefficient formula may also be appropriate. This may result in a higher similarity coefficient between those documents that include multiple occurrences of the rate or unusual concepts.

The criteria may also include distribution of duplicate concepts. That is, consideration of the number of identical or similar concepts in each document in the plurality of documents that are involved in the comparison exercise. For example, an average may be taken for each document of the number of occurrences of each concept that occurs multiple times in that document, and then an average of all of the averages is determined. Alternatively, for example, the ratio of the number of unique concepts to the total number of concepts (including duplicate concepts) throughout the documents may be determined. The resulting value may be used in the selection of the document comparison means. For example, a higher value may suggest more multiple occurrences of identical or similar concepts. Therefore, selection of document comparison means that puts less emphasis on multiple occurrences may be appropriate, such as selection of a DSS formula and/or variable measurement other than complete linkage.

Another of the criteria that may be used in selection of the document comparison means is a size distribution of the documents being compared. The "size" of the documents being considered is the number of concepts within the documents and may or may not reflect the physical size of or amount of text in the documents. The distribution of the documents may be reflected by the variance in the size of the documents. A low variance may mean that use of the Jaccard or DSS-Jaccard similarity coefficient formula may be appropriate, whereas a high variance may mean that use of the inclusion or DSS-inclusion formula may be appropriate. In case of a high variance the documents have different sizes, thus making it more likely that a large document will be compared with a small one. In this case the inclusion and DSS-inclusion formula may be preferable because they indicate if a small document is included in a large one, whereas this is less clear from the other formulas.

Thus, as indicated above, in embodiments of the invention, the comparison means for comparing two or more documents may be chosen based on criteria of the comparison. The comparison may be performed on properties of the documents that comprise, for example, numbers of concepts that are found in or are associated with the documents.

Figure 8 shows an example of a method 800 for comparing two or more documents. First, in step 802, the criteria of the comparison are determined. Examples of such criteria are given above. Next, in step 804, comparison means are selected based on the criteria. Then, in step 806, the comparison of the documents is performed using the selected comparison means. In step 808, the results are being displayed in form of a table. In step 810, the results are saved in a specific file format, such as csv. The method 800 then ends at step 812.

In alternative embodiments of the invention, a diagram may be displayed that may allow a user to visualize the similarity between two documents being compared. In some embodiments, the diagram is a two-dimensional diagram having a first axis and a second axis. Each axis corresponds to one of the documents being compared and positions along an axis indicates the positions in the corresponding document of concepts within or associated with that document.

Figure 9 shows an example of such a diagram 900. The horizontal axis corresponds to document i, whereas the vertical axis corresponds to document j. The documents i and j include the same concepts as the documents i and j shown in figures 1 to 7. The concepts within each document are shown on the corresponding axes for illustration purposes, although these may not be displayed on the diagram 900. The sequence in that the concepts are ordered on both axis corresponds to the occurance of the concepts in the documents.

Concepts that are common to both documents are highlighted on the diagram 900 at the appropriate positions on the horizontal and vertical axes with a "x", although other ways of highlighting these points are possible. Thus, in the example shown, there are 12 such points highlighted, equal to the variable c_{ij} in the complete linkage method as described above. In alternative embodiments, the principles from other methods such as integer linkage and wedding linkage may be applied to the highlighting of points on the diagram 900, possibly leading to fewer such points.

The diagram 900 is particularly useful when comparing patent documents. These documents normally have a predetermined structure and, for example, may contain one or more of the following sections in a predetermined order: background,

summary, detailed description, claims, and other sections. Therefore, where two documents are similar such that the corresponding sections include or are associated with some identical or similar concepts, the highlighted points on the diagram 900 may approximate a linear pattern. The highlighted points on the example diagram 900 could be in a generally linear arrangement as indicated by the dotted line 902, which may or may not appear on the diagram. Here, the term "linear" should be interpreted to mean that for a highlighted point, the distance from one of the axes tends to increase along with the distance from the other axis, although not necessarily in a linear manner.

Alternatively, there may occur other structures: As shown in diagram 900 all axes may be subdivided according to the document structure, meaning in several parts. Dotted line 904 divides the Y-axis, dotted line 906 divides the X-axis. In a patent document, for instance, such parts are the description and the claims. With the method described in this patent it is now possible to analyse the similarities between the parts within a document and between parts of different documents. As shown in diagram 900, a lot of the concepts both of the description and the claim part of document i is similar to a lot of the concepts of the description part of document j, but only a few to the claim part of document j. This may indicate that document j is a following document to document i. Other implications may also be obtained by this kind of analysis.

In the above description, a document is referred to as a single entity. However, a document may instead comprise multiple documents combined, or a portion of one or more documents. Where two documents are compared, this may be a comparison of two portions from the same document.

Concepts are determined in the above description for the documents being compared. However, in embodiments of the invention these concepts could be determined for a document every time the document is to be used in a comparison, or the concepts may be predetermined and retrieved when required.

Figure 10 shows an example of a data processing system 1000 that is suitable for use when implementing embodiments of the invention. The data processing system 1000

includes a central processing unit (CPU) 1002 and a main memory 1004. The system 1000 may also include a permanent storage device 1006, such as a hard disk, and/or a communications device 1008 such as a network interface controller (NIC). The system 1000 may also include a display device 1010 and/or an input device 1012 such as a mouse and/or keyboard.

The method according to the present invention may be embodied by software and/or hardwired processing means.

The documents to which the present invention is applicable may be input as linguistic data (texts), for example ASCII data in the .csv format. Inputting the data in the .csv format is particularly advantageous if the processed documents are patents which may include a different number of concepts in each patent, thus avoiding empty fields in a relational database, for example.

The output of the comparison results in accordance with the present invention may be represented by data in a database, particularly a relational database, for example in the .mdb or .assdb format. This enables a speedy processing of the obtained comparison results.

All of the features disclosed in this specification (including any accompanying claims, abstract and drawings), and/or all of the steps of any method or process so disclosed, may be combined in any combination, except combinations where at least some of such features and/or steps are mutually exclusive.

Each feature disclosed in this specification (including any accompanying claims, abstract and drawings) may be replaced by alternative features serving the same, equivalent or similar purpose, unless expressly stated otherwise. Thus, unless expressly stated otherwise, each feature disclosed is one example only of a generic series of equivalent or similar features.

Embodiments of the invention are not restricted to the details of any foregoing embodiments. Embodiments of the invention extend to any novel one, or any novel combination, of the features disclosed in this specification (including any

accompanying claims, abstract and drawings), or to any novel one, or any novel combination, of the steps of any method or process so disclosed. The claims should not be construed to cover merely the foregoing embodiments, but also any embodiments that fall within the scope of the claims.

References

The following documents are incorporated herein by reference for all purposes:

[1] J. C. GOWER, P. LEGENDRE, Metric and Euclidean properties of dissimilarity coefficients, Journal of Classification, 3 (1986) 5-48

[2] K. BACKHAUS, Multivariate Analysemethoden. Eine anwendungsorientierte Einführung, Springer, Berlin et al., 2006

[3] F. BROSIUS, SPSS 14, Redline, Heidelberg, 2006.

[4] J. QIN, Semantic Similarities between a Keyword Database and a Controlled Vocabulary Database: An Investigation in the Antibotic Resistance Literature, Journal of the American Society for Information Science, 51 (2000), 166-180

[5] R. R. BRAAM, H. F. MOED, A .F. J. VAN RAAN, Mapping of Science: Critical elaboration and new approaches, a case study in agricultural biochemistry, L. EGGHE, R. ROUSSEAU, Infometrics 87/88, Elseiver Science Publishers, Amerstdam et al., 1988

[6] P. H. A. SNEATH, R. R. SOKAL, Numerical Taxonomy, W. H. Freeman and Company, San Francisco, 1973

[7] A. DRESSLER, Patente in technologieorientierten Mergers & Acquisitions, Deutscher Universitäts-Verlag, Wiesbaden, 2006

[8] A. RIP, P. COURTAL, CO-word maps of biotechnology: An example of cognitive scientometrics, Scientometrics, 6 (1984) 381-400

[9] J. BUHWAN, D. LEE, H. CHO, J. LEE, A novel method for measuring semantic similarity for XML schema matching, Expert Systems with Applications, 34 (2008), 1651-1658

[10] V. BATAGELJ, M. BREN, Comparing Resemblance Measures, Journal of Classification, 12 (1995), 73-90

[11] A. J. TRIPPE, Patinformatics: Tasks to tools, World Patent Information, 25 (2003), 211-221

[12] J. J. SEPKOSKI, Quantified Coefficients of Association and Measurement of Similarity, Mathematical Geology, 6 (1974), 135-152

[13] L. YANHONG, T, T. RUNHUA, A Text-Mining-bases Patent Analysis in Product Innovative Process, in: N. Léon-Rvira, Trends in Computer Aided Innovation, New York, Springer-Verlag, 2007, 89-96

[14] ABOU-ASSALEH, TONY; CERCONE, NICK; KESELJ, VLADO; SWEIDAN, RAY: N-gram-based Detection of New Malicious Code, in: Proceedings of the 28th Annual International Computer Software and Applications Conference (COMPSAC'04), 2004

[15] MOENS, MARIE-FRANCINE: Information Extraction: Algorithms and Prospects in a Retrieval Context, Springer 2006

[16] KARTHIK, M. N.; DAVIS, MOSHE : Search Using N-gram Technique Based Statistical Analysis for Knowledge Extraction in Case Based Reasoning Systems CoRR cs.AI/0407009, 2004

[17] TSOURIKOV, VALERY M.; BATCHILO, LEONID S.; SOVPEL, IGOR V.: US 6,167,370. Document semantic analysis/selection with knowledge creativity capability utilizing subject-action-object (SAO) structures, 2000.

CLAIMS

1. A method of comparing first and second documents, the method comprising:

determining criteria of the comparison;

selecting comparison means based on the criteria from a plurality of comparison means; and

performing the comparison of the first and second documents by using the selected comparison means.

2. A method as claimed in claim 1, wherein using the selected comparison means comprises applying the selected comparison means to first properties of the first document and second properties of the second document.

3. A method as claimed in claim 2, wherein the comparison means includes property determining means for determining the first and second properties, and the method comprises determining the first and second properties using the property determining means.

4. A method as claimed in claim 3, wherein the properties include properties of concepts in the first and/or second documents.

5. A method as claimed in claim 4, wherein the property determining means comprises a plurality of rules each for determining a number of unique and/or repeated concepts in the first and/or second document and/or a number of concepts common to the first and second documents, and wherein selecting the comparison means comprises selecting one of the plurality of rules.

6. A method as claimed in claim 4 or 5, comprising determining the concepts in the first and/or second documents.

7. A method as claimed in any of claims 2 to 6, wherein selecting the comparison means comprises selecting at least one of a plurality of document comparison formulae each of which provide a measure of similarity of the first and second documents from the first and second properties.

8. A method as claimed in claim 7, wherein the selected at least one document comparison formula comprises at least one of a Jaccard, double-single-sided (DSS)-Jaccard, cosine, inclusion, DSS-inclusion and DSS-gamma-inclusion document comparison formulae.

9. A method as claimed in any of the preceding claims, wherein the criteria of the comparison comprise one or more of a purpose of the comparison, an importance of considering duplicate concepts in each of the first and second documents, a distribution of the duplicate concepts and a size distribution of a plurality of documents that include the first and second documents.

10. A method of comparing first and second documents, the first document including or associated with one or more first concepts, the second document including or associated with one or more second concepts, the method comprising displaying a diagram having first and second axes, the first axis corresponding to positions of the first concepts within the first document, the second axis corresponding to positions of the second concepts within the second document, the method further comprising displaying or highlighting one or more points on the diagram, each point at a position on the first axis corresponding to a first one of the concepts in the first document and on the second axis corresponding to a second one of the concepts in the second document, whereby the first concept is identical or similar to the second concept.

11. A method as claimed in claim 10, further comprising: subdividing the axes according to a common structure of the first and second documents, the common structure representative of at least two separable portions of each of the documents, thereby displaying or highlighting occurrences of concepts in one portion of the first documents identical or similar to concepts in another portion of the second document.

12. Apparatus arranged to implement the method as claimed in any of claims 1 to 11.

13. Apparatus as claimed in claim 12, wherein the apparatus comprises a data processing system.

14. A computer program comprising code for implementing a method as claimed in any of claims 1 to 11.

15. Computer readable storage storing a computer program as claimed in claim 14.

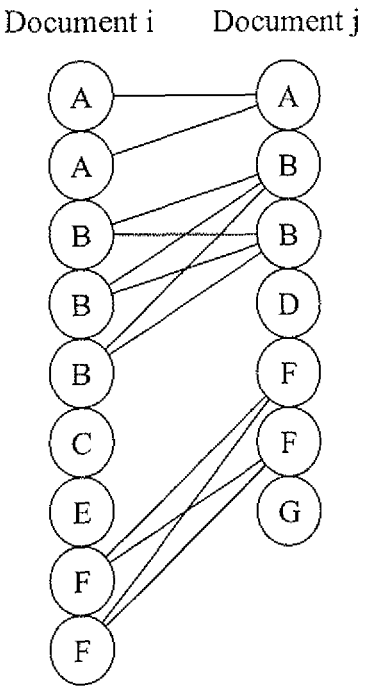


Figure 1

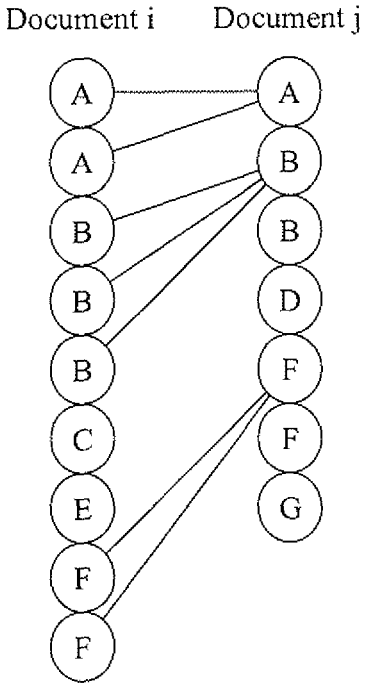


Figure 2

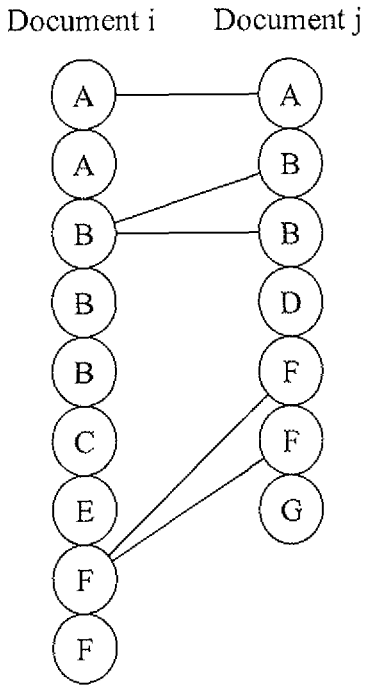


Figure 3

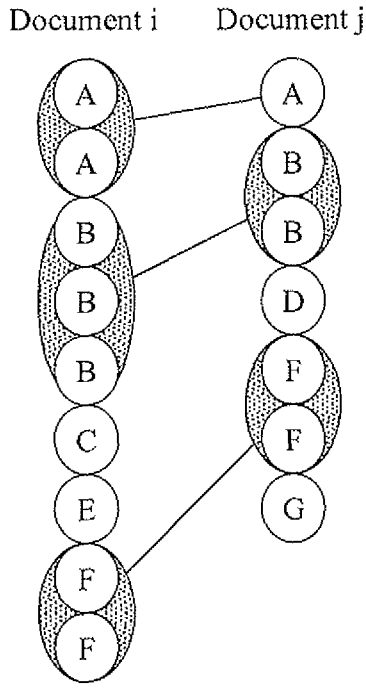


Figure 4

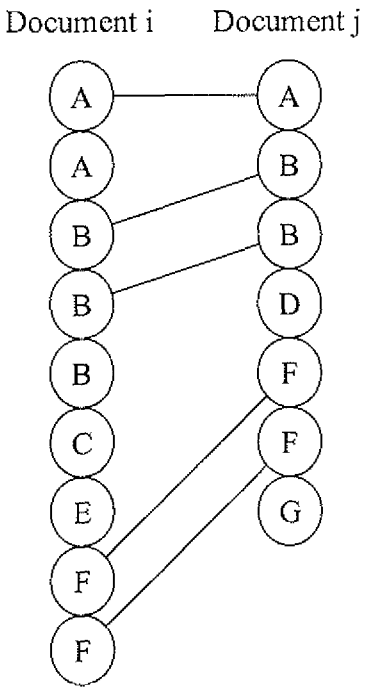


Figure 5

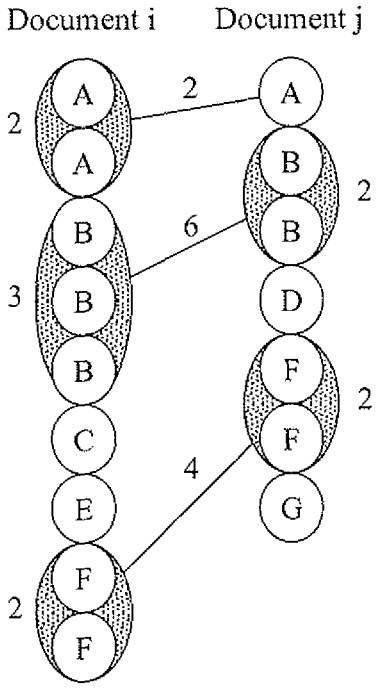


Figure 6

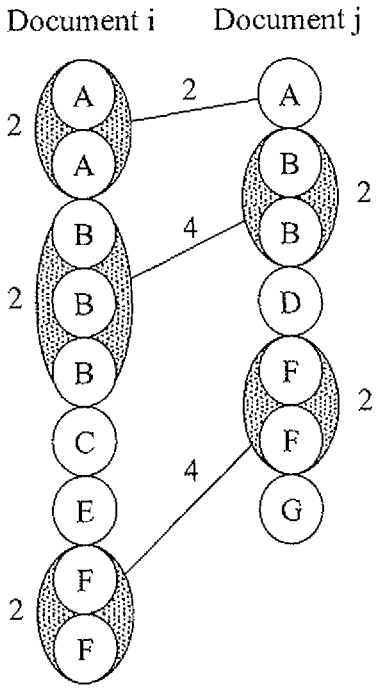
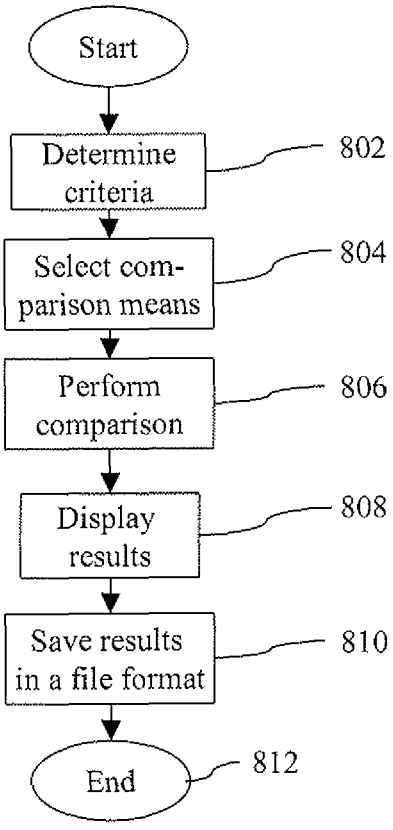
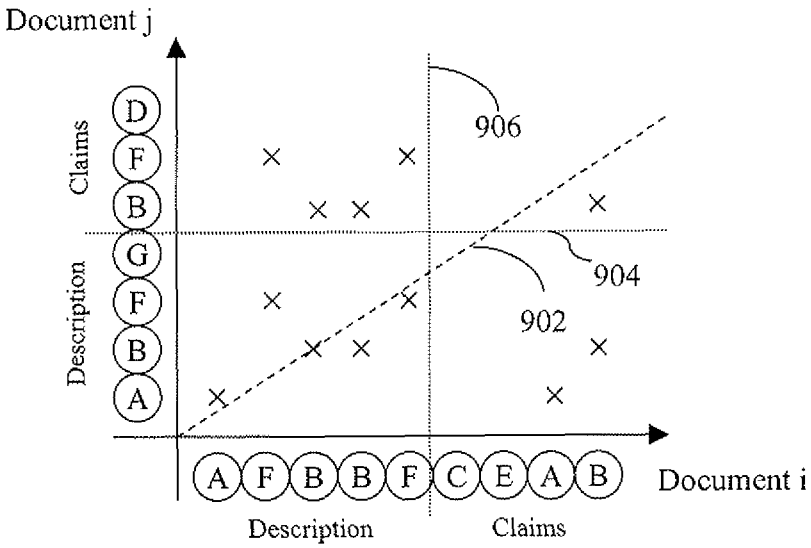


Figure 7



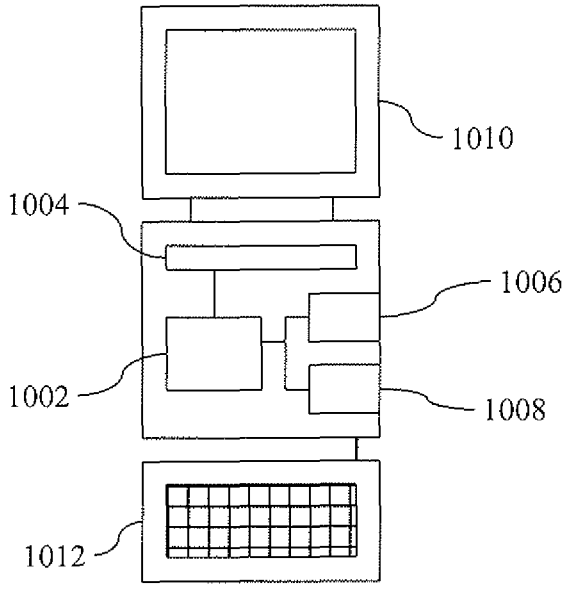
800

Figure 8



900

Figure 9



1000

Figure 10

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2009/061713

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06F17/22

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
EPO-Internal

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CHRISTIAN STERNITZKE ET AL: "Similarity measures for document mapping: A comparative study on the level of an individual scientist" SCIENTOMETRICS, KLUWER ACADEMIC PUBLISHERS, DO, vol. 78, no. 1, 20 December 2008 (2008-12-20), pages 113-130, XP019649008 ISSN: 1588-2861 abstract page 115, line 27 - page 116, line 9 -----	1-9, 12-15
X	US 2006/242140 A1 (WNEK JANUSZ [US]) 26 October 2006 (2006-10-26) abstract paragraphs [0008], [0 64] - [0065], [0130] ----- -/--	1-9, 12-15

[X] Further documents are listed in the continuation of Box C.

[X] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
1 April 2010	14/06/2010

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Lopez, Patrice

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	Dmitri Zelenko, William M. Pottenger: "Concept Space Comparison and Validation" Technical Report: UIUCDCS-R-98-2071, [Online] 11 July 1998 (1998-07-11), XP002576286 University of Illinois at Urbana-Champaign, USA Retrieved from the Internet: URL:http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.39.4075&rep=rep1&type=pdf> [retrieved on 2010-03-26] abstract page 4, lines 8-22	1-9, 12-15
X	Nico Salmaso: "SIMDISS User's Manual"[Online] December 1998 (1998-12), XP002576287 Retrieved from the Internet: URL:http://www.limno.eu/SimDiss/SDManual.pdf> [retrieved on 2010-03-26] pages 16-19	1-9, 12-15

INTERNATIONAL SEARCH REPORT

International application No.
PCT/EP2009/061713

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

see additional sheet

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying an additional fees, this Authority did not invite payment of additional fees.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

1-9(completely); 12-15(partially)

Remark on Protest

☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.

☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

☐ No protest accompanied the payment of additional search fees.

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

This International Searching Authority found multiple (groups of) inventions in this international application, as follows:

1. claims: 1-9(completely); 12-15(partially)

Method for comparing documents

---

2. claims: 10, 11(completely); 12-15(partially)

Method for visualizing differences between documents

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/EP2009/061713

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2006242140 A1	26-10-2006	NONE	