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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
26 January 2006 (26.01.2006)

PCT

(10) International Publication Number
WO 2006/008733 A3

(51) International Patent Classification:
G06E 1/00 (2006.01)

(21) International Application Number:
PCT/IL2005/000726

(22) International Filing Date: 7 July 2005 (07.07.2005)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/589,596 21 July 2004 (21.07.2004) US

(71) Applicant *(for all designated States except US)*: **EQUIVIO LTD.** [IL/IL]; 22 Hamelacha Stret, 48091 Rosh HaAyin (IL).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **MILO, Amir** [IL/IL]; 15 Hatchiya Street, 44250 Kfar Saba (IL). **RAVID, Yiftach** [IL/IL]; 23 Revivim Street, 48621 Rosh Haayin (IL).

(74) Agent: **REINHOLD COHN AND PARTNERS**; P.O.B. 4060, 61040 Tel-Aviv (IL).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SM, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, 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, HU, IE, IS, IT, LT, LU, LV, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

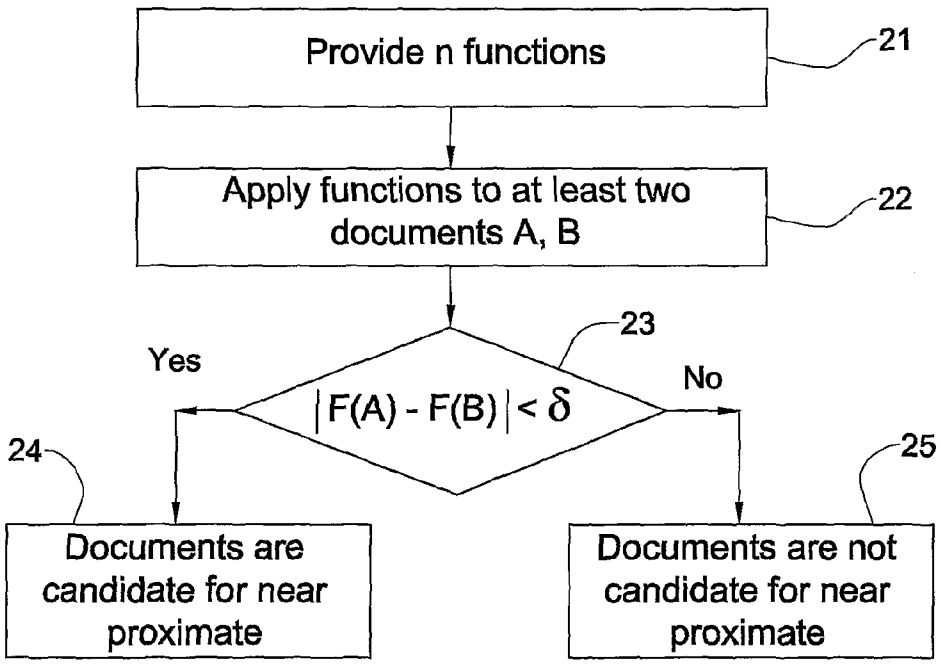
Published:
— *with international search report*
— *with amended claims*

(88) Date of publication of the international search report:
1 March 2007

Date of publication of the amended claims: 12 April 2007

[Continued on next page]

(54) Title: A METHOD FOR DETERMINING NEAR DUPLICATE DATA OBJECTS



(57) Abstract: A system for determining that a document B is a candidate for near duplicate to a document A with a given similarity level *th*. The system includes a storage for providing two different functions on the documents, each function having a numeric function value. The system further includes a processor associated with the storage and configured to determine that the document B is a candidate for near duplicate to the document A, if a condition is met. The condition includes: for any function f_i from among the two functions, f_i (A)- f_i (B) $\leq \delta_i$ (f,A,th).

WO 2006/008733 A3

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

AMENDED CLAIMS

[Received by the International Bureau on 05 February 2007(05.02.2007)]

1. A method for determining that at least one object B is a candidate for near duplicate to an object A with a given similarity level *th*, comprising

i) providing at least two different functions on an object, each function having a numeric function value;

(ii) determining that at least one object B is a candidate for near duplicate to an object A , if a condition is met, the condition includes: for at least two functions f_i from among said functions, $| f_i(A)- f_i(B) | \leq \delta_i (f,th)$, wherein δ_i is dependent upon at least f,th .

2. The method according to Claim **1**, wherein said objects being documents.

3. The method according to Claim **1**, wherein for a function f said at least one characteristics being that f is bound by a minimum value *min* and a maximum value *max,* and wherein said $\delta(f,th) = \alpha(th) \bullet | max- min |$.

4. The method according to Claim **1**, wherein for a function f said at least one characteristics being that f is not bound by a minimum value *min* and a maximum value *max*, and wherein said δ is also dependent upon A wherein $\delta(f,th,A) = \alpha(th) \bullet f(A)$.

5. The method according to Claim **2** wherein said documents include at least text and/or numbers.

6. The method according to Claim **2** wherein said documents are Microsoft office® documents.

7. The method according to Claim **2**, wherein said documents are e-mails in selected format.

8. The method according to Claim **7**, wherein said format being a member selected from a group that includes Microsoft Outlook, Lotus Notes.

9. The method according to Claim **3**, wherein $\alpha(th) = 1-th$.

10. The method according to Claim **4**, wherein $\alpha(th) = 1-th$.

11. The method according to Claim **1**, wherein at least one of said functions being a classifier.

12. The method according to Claim **11**, wherein said classifiers being of a classifier type selected from a group that includes Bayesian Classifier, Decision Trees, Support Vector Machine Classifier.

13. The method according to Claim **1**, wherein at least one of said functions is a distance function.

14. The method according to Claim **13**, wherein the providing of distance function includes: generating for each document a vector of features where each entry in the vector is the frequency/occurrence of the feature, a feature being words from the documents.

15. The method according to Claim **13**, wherein said distance function a member of a group that includes: L^{∞} (Maximum distance), L^{2} (Euclidian distance), L^{1} (sum of differences), and JS (Jensen-Shannon) distance between the two vectors.

16. The method according to Claim **2**, wherein at least one of said functions being the number of features from known type in a document.

17. The method according to any one of the preceding claims, wherein at least two of said functions are of different type.

18. The method according to Claim **1**, wherein said (i) and (ii) are applied on-line in respect of each new received object.

19. The method according to Claim **1**, further comprising: providing a database for storing signatures of objects and determining if an object has already been processed, including:

i) associating to an object a respective unique identification code;

ii) calculating a signature for the object;

iii) checking if the calculated signature is stored in the database in which case the object has already been processed; if not applying said (i) and (ii) in respect of the object and at least one other object in order to determine whether said object and at least one other object are near candidates.

20. The method according to Claim **19**, wherein said signature being checksum on an object or derivative thereof.

21. The method according to Claim **19**, wherein said database being hash table.

22. The method according to Claim **1**, further comprising applying at least one additional calculation phase in order to determine whether candidates of near duplicate objects meet a criterion for near duplicate objects.

23. The method according to Claim **22**, wherein the additional calculation phase including calculating a Resemblance between two documents.

24. The method according to Claim **22**, wherein additional calculation phase including:

iv) calculating intersection between two candidates for near duplicate objects by calculating number of shingles that are shared by the two;

v) calculating union of two candidates for near duplicate objects by calculating number of shingles that reside in either objects;

vi) determining that the two objects are near duplicate by calculating the resemblance, and in case the result exceeding a predetermined value constituting said criterion, the objects are near duplicate

25. The method according to Claim **24**, further comprising, applying an optimization for discarding candidates for near duplicate objects having a resemblance that drops below said predetermined value.

26. The method according to Claim **22**, wherein the at least one additional calculation phase is slower than the calculation of candidates of near duplicate documents, for any two documents.

27. The method according to Claim **1**, further comprising:

vii) applying at least one additional calculation phase in order to determine whether candidates of near duplicate objects meet a criterion for near duplicate objects;

viii) applying a learning phase based on objects that are determined to be candidates for near duplicate, but did not meet the criterion for near duplicate objects.

28. The method according to Claim **27**, wherein said (ii) further comprising

(1) providing additional at least one function capable of discerning between objects which were classified as candidates for near duplicate in a first phase, but did not meet said criterion in the additional phase; and

iii) applying (i) and (ii) for determining candidates for near duplicate, for any function from among said at least two functions and the additional at least one function.

29. The method according to Claim **28**, wherein at least one of said additional functions being a classifier.

30. The method according to any one of the preceding claims, further comprising: applying said (i) and (ii) in respect of more than two objects in order to determine whether at least two of said objects are near duplicate.

31. The method according to Claim **2**, for use in one or more members of the group that includes the following applications: document management, content management, digitization, legal, business intelligence, military intelligence, search engines results pre- and post-processing, archiving, source code comparisons, management of email servers, management of file servers.

32. The method according to Claim **31**, wherein said applications are marketed as a stand alone application.

33. The method according to Claim **31**, wherein said applications are marketed as (OEM).

34. A method for determining that a document A is a candidate for near duplicate to at least one other document B, comprising:

i) providing at least two different bounded functions f on document, and for each classifier providing a vector with n buckets where n is a function of th, each of size $1/n$

ii) receiving the document A, associating a unique document id to the document, and calculating a list of features;

iii) calculating a rank= f(A), where A being the list of features of the documents;

iv) add document id to buckets in the vector, as follows:

Floor(n•rank) (if greater than zero, otherwise discard this option), *Floor(n•rank)+1*, and *Floor(n•rank)+2* (if less than n , otherwise discard this option)

v) calculating union on documents id in the buckets, giving rise to set of documents id;

vi) applying (ii)-(v), in respect to a different classifier from among said at least two classifiers, giving rise to respective at least two sets of documents id;

vii) applying intersection to the at least two of the sets, stipulated in (vi), giving rise to at least two documents id, if any, being candidates for near duplicate.

35. The method according to Claim **34**, wherein said list of features being 1-grams,2-grams,3-grams,..,n-grams, where n is selected

36. The method according to Claim **34**, further comprising

applying at least one additional calculation phase in order to determine whether candidates of near duplicate objects meet a criterion for near duplicate objects.

37. The method according to Claim **36**, wherein the additional calculation phase including calculating the resemblance to verify the near-equivalence.

38. The method according to Claim **36**, wherein additional calculation phase including:

i) calculating intersection between two candidates for near duplicate objects by calculating number of shingles that are shared by the two;

ii) calculating union of two candidates for near duplicate objects by calculating number of shingles that reside in either objects;

iii) determining that the two objects are near duplicate by calculating intersection divided by union, and in case the result exceeding a predetermined value constituting said criterion, the objects are near duplicate.

39. The method according to Claim **38**, further comprising, applying an optimization for discarding candidates for near duplicate objects having a shingle ratio that drops below said predetermined value.

40. The method according to Claim **36**, wherein the at least one additional calculation phase is slower than the calculation of candidates of near duplicate documents, for any two documents.

41. The method according to Claim **1,** wherein said condition is implemented using bucket data structure.

42. A method for determining that at least one object B is a candidate for near duplicate to an object A, comprising

iv) providing at least two different functions on an object, each function having a numeric function value;

(ii) determining that at least one object B is a candidate for near duplicate to an object A , if a condition is met, the condition includes: for at least two functions f_i from among said functions, $| f_i(A)- f_i(B) | \leq \delta_i (f,A)$, wherein δ_i is dependent upon at least f and A.

43. A method for determining that at least one object B is a candidate for near duplicate to an object A, comprising

(i) providing at least two different functions on an object, each function having a numeric function value;

(ii) determining that at least one object B is a candidate for near duplicate to an object A , if a condition is met, the condition includes: for at least two functions f_i from among said functions a relationship between results of the functions when applied to the objects meets a given score.

44. The method according to Claim **43**, wherein said relationship being $| f_i(A)- f_i(B) |$, and said score being $\delta_i (f,A)$, wherein δ_i is dependent upon at least f and A, and wherein said condition is met if $| f_i(A)- f_i(B) | \leq \delta_i (f,A)$.

45. A system for determining that at least one object B is a candidate for near duplicate to an object A, comprising:

a storage providing at least two different functions on an object, each function having a numeric function value;

a processor associated with said storage and configured to determine that at least one object B is a candidate for near duplicate to an object A, if a condition is met, the condition includes: for at least two functions f_i from among said functions, $| f_i(A)- f_i(B) | \leq \delta_i (f,A)$, wherein δ_i is dependent upon at least *f* and *A*.

46. The system according to Claim **45**, wherein said determining that at least one object B is a candidate for near duplicate to an object A with a given similarity level *th*, and wherein said δ_i is further dependent upon *th*.

47. A system for determining that at least one object B is a candidate for near duplicate to an object A, comprising: a storage providing at least two different functions on an object, each function having a numeric function value;

a processor associated with said storage, configured to determine that at least one object B is a candidate for near duplicate to an object A , if a condition is met, the condition includes: for at least two functions f_i from among said functions, a relationship between results of the functions when applied to the objects meets a given score.

48. A computer product comprising a storage for storing computer code portions capable of performing the method stages according to Claim **1**.

49. A computer product comprising a storage for storing computer code portions capable of performing the method stages according to Claim **34**.

50. A computer product comprising a storage for storing computer code portions capable of performing the method stages according to Claim **42**.

51. A computer product comprising a storage for storing computer code portions capable of performing the method stages according to Claim **43**.

52. The method according to Claim **1**, wherein of said objects being voice data, and further comprising, converting said objects to respective text based documents.

53. The method according to Claim **34**, wherein of said objects being voice data, and further comprising, converting said objects to respective text based documents.

54. The method according to Claim **42**, wherein of said objects being voice data, and further comprising, converting said objects to respective text based documents.

55. The method according to Claim **43**, wherein of said objects being voice data, and further comprising, converting said objects to respective text based documents.

56. A method for determining that at least one object B is a candidate for near duplicate to an object A with a given similarity level, comprising

i) calculating at least two different numeric values for each object;

ii) determining that an object B, from among said objects, is a candidate for near duplicate to the object A, if a respective condition is met for at least two pairs of values, where each pair includes a distinct numeric value from among the values of object B and a corresponding distinct numeric value from among the values of object A.

57. The method according to Claim **56**, wherein said condition includes a condition regarding the distance between the numeric values of the pair.

58. The method according to Claim **56**, wherein said objects being documents.

59. The method according to Claim **57**, wherein said objects being documents.

60. A system for determining that at least one object B is a candidate for near duplicate to an object A with a given similarity level, comprising: a processor associated with a storage, configured to perform at least the following:

iii) calculating at least two different numeric values for each object;

(ii) determining that an object B, from among said objects, is a candidate for near duplicate to the object A, if a respective condition is met for at least two pairs of values, where each pair includes a distinct numeric value from among the values of object B and a corresponding distinct numeric value from among the values of object A.

61. The system according to Claim **60**, wherein said condition includes a condition regarding the distance between the numeric values of the pair.

62. The system according to Claim **60**, wherein said objects being documents.

63. The system according to Claim **61**, wherein said objects being documents.