(12) DEMANDE INTERNATIONALE PUBLIÉE EN VERTU DU TRAITÉ DE COOPÉRATION EN MATIÈRE DE BREVETS (PCT)

(19) Organisation Mondiale de la Propriété Intellectuelle
Bureau international

(43) Date de la publication internationale
23 janvier 2014 (23.01.2014)

WIPO | PCT

(10) Numéro de publication internationale
WO 2014/013191 A2

(51) Classification internationale des brevets : *G06F 17/30* (2006.01)

(21) Numéro de la demande internationale : PCT/FR2013/051713

(22) Date de dépôt international : 16 juillet 2013 (16.07.2013)

(25) Langue de dépôt : français

(26) Langue de publication : français

(30) Données relatives à la priorité :
1256922 17 juillet 2012 (17.07.2012) FR

(71) Déposants : **UNIVERSITE LILLE 1 - SCIENCES ET TECHNOLOGIES** [FR/FR]; Cité Scientifique, F-59655 Villeneuve d'Ascq (FR). **CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE - CNRS** [FR/FR]; 3 rue Michel-Ange, F-75794 Paris (FR).

(72) Inventeurs : **LANCIERI, Luigi**; 188 bis rue des Patriotes, F-59150 Wattrelos (FR). **LEPRETRE, Eric**; 38 rue du Pévèle, F-59242 Templeuve (FR).

(74) Mandataire : **BETHENOD, Marc**; Novagraaf Technologies, 122 rue Edouard Vaillant, F-92593 Levallois-Perret (FR).

(81) États désignés *(sauf indication contraire, pour tout titre de protection nationale disponible)* : AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, 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, 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, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) États désignés *(sauf indication contraire, pour tout titre de protection régionale disponible)* : ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), eurasien (AM, AZ, BY, KG, KZ, RU, TJ, TM), européen (AL, 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, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Publiée :

— *sans rapport de recherche internationale, sera republiée dès réception de ce rapport (règle 48.2.g))*

(54) **Title** : METHOD OF MEASURING AN INDEX CHARACTERISING THE CONTENT OF AN ELECTRONIC DOCUMENT

(54) **Titre** : PROCEDE DE MESURE D'UN INDICE CARACTERISANT LE CONTENU D'UN DOCUMENT NUMERIQUE

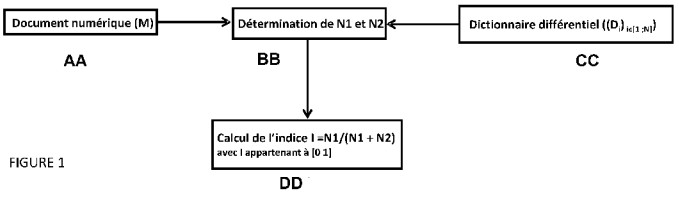

FIGURE 1

AA Electronic document (M)
BB Determination of N1 and N2
CC Differential dictionary $((D_i)_{ie[1;N]})$
DD Calculation of the index I=N1/(N1 + N2) with I belonging to [0 1]

(57) **Abstract** : The invention relates to a method for measuring an index (I) characterising the content of an electronic document (M) comprising descriptive elements from a differential dictionary $((D_i)_{ie[1;N]}$ comprising two sets U_1 and U_2 of criteria, said criteria relating to elements which have a format similar to that of the descriptive elements and each being associated with a relevance factor (HK), said method comprising the following steps: - application of the criteria of said sets U_1 and U_2 to the descriptive elements of the content of said electronic document (M) in such a way as to determine the numbers N1 and N2 of the descriptive elements of said content corresponding to the criteria of said sets, U_1 and U_2 respectively, and - calculation of the index (I) included in the range of values [0,1] by the execution of the operation: N1/(N1 + N2).

(57) **Abrégé** : L'invention concerne

[Suite sur la page suivante]

WO 2014/013191 A2

un procédé de mesure d'un indice (I) caractérisant le contenu d'un document numérique (M) comportant des éléments descriptifs à partir d'un dictionnaire différentiel ($(D_i)_{i\in[1;N]}$) comprenant deux ensembles U_1 et U_2 de critères, lesdits critères se rapportant à des éléments ayant un format similaire à celui des éléments descriptifs et étant chacun associé à un coefficient de pertinence (HK), ledit procédé comportant les étapes suivantes : - application des critères desdits ensembles U1 et U_2 aux éléments descriptifs du contenu dudit document numérique (M) de sorte à déterminer les nombres N1 et N2 d'éléments descriptifs dudit contenu correspondant aux critères desdits ensembles respectivement U_1 et U_2, et -calcul de l'indice (I) compris dans l'intervalle de valeurs [0,1] par l'exécution de l'opération : N1/(N1 + N2).

PROCEDE DE MESURE D'UN INDICE CARACTERISANT LE CONTENU D'UN DOCUMENT NUMERIQUE

DOMAINE TECHNIQUE DE L'INVENTION

[001] La présente invention concerne les procédés d'évaluation de la pertinence d'un document numérique. Plus particulièrement, l'invention concerne un procédé de mesure d'un indice caractérisant le contenu d'un document numérique. Un tel indice étant apte à décrire le contenu d'un tel document.

ÉTAT DE LA TECHNIQUE

[002] Dans les domaines relatifs au classement de documents numériques, ou encore de tri de tels documents en vu de leur utilisation dans le cadre de moteur de recherche par exemple, il y a un besoin croissant de procédé permettant de mesurer la pertinence de tels documents.

[003] La nature de ces documents numériques, textes d'opinion traitant de sujets variés, et l'identification plus précise de l'objet de tel document, requièrent une définition des outils d'analyse qui soient plus fine et donc susceptibles de permettre une identification plus précise quant au sens du contenu de tels documents, par exemple par l'extraction et la quantification de sentiments qui peuvent y être exprimés.

[004] On connaît par exemple dans l'état de l'art, des procédés qui prévoient le classement de documents numériques à partir de la détermination de la teneur des termes mentionnés dans le contenu de ces documents. Pour ce faire de tels procédés prévoient l'utilisation d'outils d'analyse capable d'extraire et de quantifier les éléments du contenu de ce document numérique.

Dans de tels procédés, des lexiques de termes ou d'expressions sont mis en œuvre afin de compléter le traitement du contenu de ce document numérique par une analyse sémantique des termes qui le constituent.

Chaque document numérique, dans ces procédés, est alors considéré comme un ensemble de mots (BOW acronyme de « Bag Of Words ») représenté par un vecteur de caractéristiques, vecteurs classés ensuite selon des techniques d'apprentissage de type réseau bayésien, par exemple un modèle de classification Naïve Bayes(NB), d'entropie maximale (ME Maximum Entropy) ou encore de Support Vector Machine (SVM).

De plus ces procédés mettent en œuvre des processus de traitement pour la prise en compte de l'ordre des mots constituant ces textes, des structures syntaxiques et des relations sémantiques entre les mots comme par exemple : POS (Part Of Speech) reposant sur le rôle des mots dans les phrases, WR (Word Relation) n-gram pour les relations de dépendances entre les mots, TF-IDF (TermFrequency - Inverse Document Frequency) pour la pondération en fonction de la fréquence de chaque mot.

[005] Cependant un inconvénient majeur de tels procédés réside dans le fait que lors de la phase d'apprentissage, il est nécessaire qu'une intervention humaine conséquente soit réalisée lors de l'évaluation des textes de références.

On comprendra que de telles interventions humaines ont pour inconvénients principaux de prendre du temps et de nécessiter des compétences spécifiques avec un haut niveau de technicité. Ces contraintes ont pour conséquences de fournir des résultats ayant une portée limitée en termes de généricité (une langue, un seul contexte, ..).

[006] La présente invention vise alors résoudre ces problèmes résultant des inconvénients de l'état de l'art.

DIVULGATION DE L'INVENTION

[007] Un problème qui se pose est donc d'améliorer les performances de procédés de mesure d'un indice caractérisant le contenu d'un document numérique.

[008] La présente invention se propose d'automatiser, et d'optimiser les procédés de mesure d'un indice caractérisant le contenu d'un document numérique.

[009] L'invention permet d'augmenter la généricité et de réduire le temps de traitement mis en œuvre par un tel procédé qui réduit notablement le besoin d'intervention humaine, en particulier, pour la constitution du dictionnaire.

[0010] L'invention requiert un coût de mis en œuvre réduit par rapport aux procédés de l'état de l'art sans conséquence sur son efficacité.

[0011] De plus, l'invention reste indépendante de la langue utilisé dans les documents numériques devant faire l'objet d'une mesure d'indice caractérisant le contenu d'un document numérique.

[0012] Dans ce dessein, un aspect de l'invention se rapporte à procédé de mesure d'un indice (I) caractérisant le contenu d'un document numérique (M) comportant des éléments descriptifs à partir d'un dictionnaire différentiel $((D_i)_{i\epsilon[1\ ;N]})$ comprenant deux ensembles U_1 et U_2 de critères, lesdits critères se rapportant à des éléments ayant un format similaire à celui des éléments descriptifs de (M) et étant chacun associé à un coefficient de pertinence (HK), ledit procédé comportant les étapes suivantes:

- application des critères desdits ensembles U_1 et U_2 aux éléments descriptifs du contenu dudit document numérique (M) de sorte à déterminer les nombres N1 et N2 d'éléments descriptifs dudit contenu correspondant aux critères desdits ensembles respectivement U_1 et U_2, et

- calcul de l'indice (I) compris dans un intervalle de valeurs [0,1] par l'exécution de l'opération : N1/(N1 + N2).

[0013] Selon des modes de réalisation particuliers :

- l'ensemble (U1) comporte des critères de même catégorie $(((C_i)_{i\epsilon[1\ ;N]})+HK)$ avec HK>0 et ledit ensemble (U2) comporte des critères de même catégorie $(((C_j)_{j\epsilon[1\ ;N]}) + HK)$ avec HK<0 ;

- le dictionnaire différentiel $((D_i)_{i\epsilon[1;N]})$ est généré à partir des étapes suivantes :

  - comparaison entre deux documents numériques sources (S1,S2), comprenant chacun une liste d'éléments associés avec respectivement un indice I_1 et un indice I_2 de redondance ;

  - conception de critères ;

- l'étape de comparaison comporte l'une des sous-étapes suivantes :

  - identification d'un élément compris dans un seul des documents numériques sources (S1,S2), ou

  - identification d'un élément compris dans les deux documents numériques sources (S1,S2), et que son indice (I_1, I_2) dans l'un de ces deux documents soit x fois supérieure à son indice (I_1, I_2) dans l'autre de ces deux documents ;

- la génération du dictionnaire différentiel $((D_i)_{i\epsilon[1;N]})$ comporte une étape de conception de critères par l'association à l'élément identifié lors de l'étape de comparaison d'un coefficient de pertinence HK correspondant à la différence potentiellement négative entre les indices I1 et I2 ;

- les deux documents numériques sources (S1,S2) sont générés lors des étapes suivantes :

  - concaténation de documents numériques initiaux (C) en deux documents numériques distincts $((A_i)_{i\epsilon[1;N]})$ et $((B_i)_{i\epsilon[1;N]})$, et

  - classement des éléments contenus dans chacun de ces documents numériques distincts $((A_i)_{i\epsilon[1;N]})$ et $((B_i)_{i\epsilon[1;N]})$ en fonction de leurs indices de redondances (I_1,I_2) ;

- les documents numériques initiaux (C) sont extraits de base de données à partir de requêtes comportant des termes d'amorces dont les caractéristiques sémantiques sont opposables ;

- les documents numériques initiaux (C) sont extraits de base de données sans l'utilisation de requêtes comportant des termes d'amorces mais en fonction des critères dits de collectes desdits documents numériques initiaux (C) ;

- le dictionnaire $((D_i)_{i\in[1;N]})$ généré sur la base des documents numériques distincts $((A_i)_{i\in[1;N]})$ et $((B_i)_{i\in[1;N]})$ est appliqué aux documents numériques initiaux C de sorte à les classer en fonction de leur indice (I) et à générer de nouveaux documents numériques distincts $((A_i)_{i\in[1;N]})$ et $((B_i)_{i\in[1;N]})$;

- Il comprend la génération d'un dictionnaire $((D_i)_{i\in[1;N]})$ à partir des nouveaux documents numériques distincts $((A_i)_{i\in[1;N]})$ et $((B_i)_{i\in[1;N]})$, et

- un dictionnaire synthétique (S) est réalisé à partir de plusieurs dictionnaires $((D_i)_{i\in[1;N]})$.

DESCRIPTION DES FIGURES

D'autres avantages et caractéristiques de l'invention apparaîtront mieux à la lecture de la description d'un mode de réalisation préféré qui va suivre, en référence à la figure 1, réalisé à titre d'exemple indicatif et non limitatif :

- figure 1 représente un organigramme relatif au procédé selon un mode de réalisation de l'invention ;

- figure 2 représente un organigramme relatif à la génération d'un dictionnaire différentiel $((D_i)_{i\in[1;N]}$ selon un mode de réalisation de l'invention,;

- figure 3 représente un organigramme plus détaillé portant sur la génération d'un dictionnaire différentiel $((D_i)_{i\in[1;N]}$ selon un mode de réalisation de l'invention, et

- figure 4 représente un organigramme portant sur la génération d'un dictionnaire différentiel $((D_i)_{i\in[1;N]}$ selon un autre mode de réalisation de l'invention.

MODES DE REALISATION DE L'INVENTION.

[0014] Dans un mode de réalisation de l'invention le procédé de mesure d'un indice I caractérisant le contenu d'un document numérique M est mis en œuvre par un dispositif informatique.

[0015] De manière générale, ce dispositif informatique correspond, de manière non limitative, à tous dispositifs comprenant :

- des moyens de traitement : au moins un microprocesseur, et des moyens de mémoire (mémoire volatile et/ou non volatile et/ou de masse) ;

- des moyens de saisie, tel qu'un clavier et/ou une souris et/ou écran tactile, ou encore à des moyens de commande vocale ;

- des moyens d'affichage ;

- des moyens de communication ;

Les moyens de communication de ce dispositif se rapportent par exemple aux technologies et/ou normes suivantes :

- WI-FI (abréviation de wirelessfidelity) et/ou Wimax, et GPRS (General Packet Radio Service), GSM, UMTS, HSDPA ou IMS (IP MultimediaSubsystem), ou

- Ethernet.

[0016] On notera, en particulier, que ces moyens de communication sont compatibles avec les moyens de communication du serveur mettant en œuvre les bases de données.

[0017] Les moyens de traitement de ce dispositif informatique sont aptes à mettre en œuvre un code informatique se rapportant à un algorithme correspondant à un programme d'ordinateur, archivé dans les moyens de mémoire du dispositif informatique, comportant des instructions aptes à effectuer les traitements nécessaires pour réaliser les étapes du procédé selon un mode de réalisation de l'invention.

[0018] Ainsi qu'il est illustré à la figure 1, le procédé de mesure d'un indice I caractérisant le contenu d'un document numérique M est réalisé à partir notamment d'un dictionnaire différentiel $((D_i)_{i\in[1;N]})$.

[0019] Cet indice I dans ce mode de réalisation de l'invention permet de mesurer la valence émotionnelle (par exemple le sentiment positif ou négatif) des éléments descriptifs constitutifs du contenu du document numérique M que sont par exemple les expressions textuelles, les termes, les expressions du langage ou les mots. A titre d'exemple les éléments constitutifs du contenu du document numérique M peuvent se rapporter à un ensemble de caractères ou encore à des symboles graphiques représentant un phonème ou une syllabe, ou encore une ou plusieurs unités de sens (par exemple un idéogramme).

Dans les modes de réalisation décrits, ces éléments constitutifs du contenu de document numérique sont de manière non limitative des termes.

[0020] Le document numérique M correspond à tous les types ou formats de fichiers numériques dans la mesure où leur contenu peut être exploité par ledit procédé. Dans le présent mode de réalisation, ce document numérique M est de manière non limitative, un document textuel soit un texte.

[0021] Si par exemple, ce document numérique M est un document audio ou vidéo, le contenu de celui-ci peut alors être converti en symboles graphiques (e.g texte), selon des technologies de conversion bien connues de l'état de l'art, afin d'être exploitable par le procédé selon un mode de réalisation de l'invention.

[0022] Le dictionnaire différentiel $((D_i)_{i\in[1;N]})$ comprend deux ensembles U1 et U2 de critères. Ces critères permettent de mesurer cet indice en déterminant la valence émotionnelle des éléments constituant le contenu du document numérique.

[0023] Chacun des deux ensembles de critères U1 et U2 comprend par exemple pour l'ensemble U1 de critères, des éléments se rapportant à des termes connotés positivement, et pour l'ensemble U2 des éléments se

rapportant à des termes connotés négativement. Plus précisément, ces critères correspondent à une association de l'élément, ici le terme, avec un coefficient de pertinence : HK. Cet élément pour U1 est $((C_i)_{i\epsilon[1\ ;N]})$ et pour U2 $((C_j)_{j\epsilon[1\ ;N]})$. Ces éléments $((C_i)_{i\epsilon[1\ ;N]})$ et $((C_j)_{j\epsilon[1\ ;N]})$ sont du même format que ceux qui constituent le document numérique M. Si ce document numérique M est constitué d'idéogrammes alors les éléments $((C_i)_{i\epsilon[1\ ;N]})$ et $((C_j)_{j\epsilon[1\ ;N]})$ seront des idéogrammes. Si comme dans le présent mode de réalisation de l'invention le document numérique M est constitué de termes alors les éléments $((C_i)_{i\epsilon[1\ ;N]})$ et $((C_j)_{j\epsilon[1\ ;N]})$ seront des termes.

HK correspond à la différence potentiellement négative entre les indices I1 et I2. HK est une fréquence différentielle. Soit HK= I1-I2.

Les indices de redondances I1 et I2 sont définis plus en détail par la suite, et correspondent à la fréquence d'apparitions de ce terme dans respectivement chacun des documents numériques distincts respectivement $((A_i)_{i\epsilon[1\ ;N]})$ et $((B_i)_{i\epsilon[1\ ;N]})$. Autrement dit pour un terme donné présent à la fois dans les documents numériques distincts respectivement $((A_i)_{i\epsilon[1\ ;N]})$ et $((B_i)_{i\epsilon[1\ ;N]})$ on considérera par exemple que la redondance de ce terme dans le document $((A_i)_{i\epsilon[1\ ;N]})$ est I1 et sa redondance dans le document $((B_i)_{i\epsilon[1\ ;N]})$ est I2. Si le terme est présent 2 fois dans le document $((A_i)_{i\epsilon[1\ ;N]})$ et absent du document $((B_i)_{i\epsilon[1\ ;N]})$ alors I1=2 et I2=0. Les indices de redondances I1 et I2 selon le même raisonnement sont également utilisés dans les documents numériques sources (S1,S2).

Par conséquent, compte tenu du fait que $((A_i)_{i\epsilon[1\ ;N]})$ et $((B_i)_{i\epsilon[1\ ;N]})$ sont de tailles comparables, HK représente donc la différence de fréquence (autrement appelé fréquence différentielle) d'apparition du terme entre les documents numériques distincts $((A_i)_{i\epsilon[1\ ;N]})$ et $((B_i)_{i\epsilon[1\ ;N]})$.

Ces documents numériques distincts respectivement $((A_i)_{i\epsilon[1\ ;N]})$ et $((B_i)_{i\epsilon[1\ ;N]})$ correspondent par exemple à du texte.

[0024] On comprend bien que l'ensemble :

- U1 comporte des critères de même catégorie : $((C_i)_{i\epsilon[1;N]})$ + HK (avec HK>0)

- U2 comporte des critères de même catégorie : $((C_j)_{j\epsilon[1;N]})$ + HK (avec HK<0)

[0025] Ce dictionnaire différentiel $((D_i)_{i\epsilon[1;N]})$ est utilisé pour mesurer l'indice I caractérisant le contenu du document numérique M. Le dictionnaire différentiel $((D_i)_{i\epsilon[1;N]})$ permet au regard de la méthode par laquelle il est généré d'apporter une précision optimale dans la mesure de cet indice I. Cette précision est telle que le procédé permet par exemple de déterminer le sentiment émotionnel qui émane d'un document numérique M tel que la peur, la confiance, etc...

[0026] Ce procédé comprend une étape d'application de ces critères aux éléments du contenu dudit document numérique M de sorte à déterminer le nombre d'élément N1 et N2 du contenu du document numérique M correspondant respectivement à un des critères desdits ensembles U1 et U2.

Cette étape consiste à vérifier dans le dictionnaire $((Di)_{i\epsilon[1;N]})$ la présence de chaque élément de N. Si l'élément est présent assorti d'une valeur HK positive alors l'élément est considéré appartenir à l'ensemble U1. Si il est présent assortit d'une valeur HK négative, il est considéré appartenir à l'ensemble U2.

Par exemple, lorsque le document numérique M correspond à un document textuel provenant par exemple d'Internet, ou d'un contenu d'un courrier électronique, etc... La mesure de l'indice I, caractérisant le contenu de ce document textuel, est réalisée de la manière suivante :

- détermination du nombre d'éléments N1 ici des termes de ce document textuel correspondant aux critères compris dans l'ensemble U1 ;

- détermination du nombre d'éléments N2 ici des termes de ce document textuel correspondant aux critères compris dans l'ensemble U2.

[0027] Une fois les critères appliqués aux éléments du contenu de ce document numérique M et les nombres N1 et N2 obtenus, une valeur de cet

indice I est calculée par l'exécution d'une opération par les moyens de traitement du dispositif informatique. Cette opération se rapporte alors à :

$I = N1/ (N1 + N2)$.

[0028] Une fois la valeur de cet indice I déterminée le document numérique M est archivé dans une base de données avec la valeur de cet indice et le domaine auquel ce document se rapporte. Ce document numérique M est alors classé dans cette base de données de documents numériques en fonction de la valeur de son indice I.

[0029] Par conséquent, la valeur de I évoluant entre une valeur de zéro et de un représente le niveau du sentiment, d'autant plus positif s'il est proche de un et d'autant plus négatif s'il est proche de zéro. Le type de sentiment (confiance, satisfaction, peur, ..) ainsi que le domaine (Restauration, achat informatique, politique, ..) est induit par le contenu du dictionnaire $((Di)_{i\epsilon[1\ ;N])}$.

[0030] Ainsi qu'il est illustré aux figures 2 et 3, ce dictionnaire est généré à partir notamment de deux documents numériques distincts $((A_i)_{i\epsilon[1\ ;N]})$ et $((B_i)_{i\epsilon[1\ ;N]})$, lesquelles peuvent être obtenus à partir de deux modes de réalisation différents : un mode dit supervisé et un autre mode dit non supervisé.

[0031] On notera que l'approche non supervisée est la plus performante compte tenu du besoin très limité d'intervention humaine.

Cependant l'approche supervisée garde de son intérêt car elle permet un contrôle plus important de l'humain sur la constitution du dictionnaire différentiel et d'aborder plus de type de sentiments que dans le mode de réalisation non supervisé.

[0032] Dans le premier mode de réalisation dit supervisé, le dictionnaire différentiel $((D_i)_{i\epsilon[1\ ;N]})$ mis en œuvre dans le procédé de mesure de l'indice I est généré à partir de :

- une étape de comparaison entre deux documents numériques sources S1 et S2, comprenant chacun une liste d'éléments associés avec respectivement un indice I1 et un indice I2 de redondance et

- une étape de conception de critères.

[0033] Dans ce mode de réalisation, ces deux documents sources S1 et S2 se rapportent à un même domaine technique ou contexte, comme par exemple la restauration ou l'informatique, et sont générés par le dispositif informatique à partir dans un premier temps de mots clés, encore appelés mot clés d'amorces ou termes d'amorces. Ces mots clés d'amorces sont fournis par un opérateur.

Le document source S1 comporte des éléments qui définissent ce domaine de manière différente de ceux compris dans le document source S2.

Ces mots clés d'amorces caractérisent un domaine donné en correspondant à des termes exprimant un sentiment ou une opinion, ou encore des caractéristiques de ce domaine.

Par exemple ces mots clés d'amorces peuvent être des termes visant à évaluer :

- la réputation de restaurants comme par exemple : succulent, chaleureux (termes positifs) ; froid, mauvais (termes négatifs);

- la réputation d'un vendeur de matériel informatique : fiable, compétent (termes positifs); pannes, incompétent (termes négatifs) ;

- le sentiment de sécurité perçu : confiance, réticence.

Ces termes d'amorces peuvent être renseignés par un opérateur avec pour contrainte essentielle que sur la caractéristique sémantique ces termes puissent être opposés. Par exemple ces mots clés définissant les caractéristiques du domaine peuvent correspondre à un sentiment ou une opinion positive et négative. Pour la restauration : succulent (terme positif) ; mauvais (terme négatif). On dit ici que les caractéristiques sémantiques de ces mots clefs sont opposables.

[0034] Dans le cadre de la génération de ce dictionnaire différentiel $((D_i)_{i\in[1;N]})$, de manière générale deux à trois mots clés d'amorces par catégories (positif ou négatif) sont nécessaires.

[0035] A partir de ces mots clés, le dispositif informatique génère une requête de manière à réaliser une recherche dans la base de données, ou de manière alternative dans d'autres bases de données ou encore Internet. Cette recherche vise alors à retrouver des documents numériques initiaux C dans chacune des catégories - positif, négatif – à partir d'une requête comprenant des mots clefs d'amorces positifs et une autre requête comprenant des mots clefs d'amorces négatifs. Les documents numériques initiaux C sont par la suite convertis au format de documents textuels contenant ces termes si ces documents numériques initiaux C sont dans des formats différents.

[0036] Ces documents numériques initiaux C sont ensuite concaténés en deux documents numériques distincts $((A_i)_{i\in[1;N]})$ et $((B_i)_{i\in[1;N]})$ qui correspondent chacun à une catégorie positive ou négative.

[0037] La contrainte ici est de recueillir un nombre de documents numériques initiaux C de tailles proches en nombre suffisant (10 au minimum). On peut toutefois tolérer une certaine dispersion : par exemple, 10 % des documents numériques initiaux C pourront être de 2 à 3 fois la taille moyenne des documents numériques initiaux C constituant le résultat de la recherche. Le nombre nécessaire est fixé par la contrainte induite par la taille des documents A et B de l'ordre de 100 à 150 ko par fichier correspondant à un document.

Cette taille est suffisante. Les performances s'améliorent très peu avec des tailles plus importantes. En effet, si les tailles sont trop faibles la proportion de termes « inutiles » pourra être très différente d'un document numérique distinct à l'autre. En augmentant la taille, la fréquence des termes inutiles augmente et se stabilise à un niveau identique dans les deux documents numériques distincts. Aller au delà de cette taille n'apporte rien dans la performance du processus de discrimination.

[0038] Ces deux documents numériques distincts $((A_i)_{i\epsilon[1\ ;N]})$ et $((B_i)_{i\epsilon[1\ ;N]})$ se rapportent à des fichiers qui sont de manière générale de taille similaire.

Ces documents numériques distincts $((A_i)_{i\epsilon[1\ ;N]})$ et $((B_i)_{i\epsilon[1\ ;N]})$ font par la suite l'objet d'un traitement particulier lors d'une étape d'analyse fréquentielle afin d'identifier les termes de leur contenu et leur fréquence d'apparitions, soit leur indice de redondance I1 et I2.

En effet, ces deux documents numériques distincts $((A_i)_{i\epsilon[1\ ;N]})$ et $((B_i)_{i\epsilon[1\ ;N]})$ contiennent des termes connotés qu'il convient d'identifier. On sait simplement que l'un contient plutôt des termes négatifs et l'autre des termes positifs. Dans ce mode de réalisation de l'invention, les termes d'amorces ont été choisis en conséquence. Donc, chacun des documents numériques distincts $((A_i)_{i\epsilon[1\ ;N]})$ et $((B_i)_{i\epsilon[1\ ;N]})$ contient des termes utiles (termes connotés négativement ou positivement) et des termes inutiles (articles, mots non connotés,..). Les termes inutiles sont statistiquement autant représentés dans les deux documents numériques distincts $((A_i)_{i\epsilon[1\ ;N]})$ et $((B_i)_{i\epsilon[1\ ;N]})$. En considérant par exemple que ces documents numériques distincts $((A_i)_{i\epsilon[1\ ;N]})$ et $((B_i)_{i\epsilon[1\ ;N]})$ sont relatifs à des fichiers ayant une taille « suffisante » et similaire (voir plus loin étape de vérification de cohérence), on détecte alors cette équivalence et les termes inutiles sont alors supprimés. L'avantage relatif à la conservation des termes utiles et l'élimination des autres est d'autant meilleur que les listes sont de tailles proches.

La suppression des termes inutiles est réalisée automatiquement car pour qu'un terme soit retenu dans le cadre de la génération d'un dictionnaire il faut, que lors d'une étape de sélection, que ce terme soit compris au moins dans un des deux documents $((A_i)_{i\epsilon[1\ ;N]})$ et $((B_i)_{i\epsilon[1\ ;N]})$ et que son indice de redondance (I1, I2) dans l'un des deux documents $((A_i)_{i\epsilon[1\ ;N]})$ et $((B_i)_{i\epsilon[1\ ;N]})$ soit x fois supérieure à son indice redondance (I1, I2) dans l'autre des deux documents $((A_i)_{i\epsilon[1\ ;N]})$ et $((B_i)_{i\epsilon[1\ ;N]})$. Par exemple x=2.

Soit $I1 > 2 \times I2$ ou $I2 > 2 \times I1$.

Du fait que les termes inutiles ont des fréquences d'apparition proches quelque soit la polarité du sentiment exprimés (Ai, Bi). Ils ne satisferont

jamais les contraintes suivantes : I1 > 2 x I2 ou I2 > 2 x I1. Et donc ces termes ne seront pas conservés dans le dictionnaire différentiel.

[0039] Lors d'une étape de vérification de la cohérence des fichiers A et B, le contenu de ces deux fichiers fait l'objet d'un traitement particulier visant à contrôler que l'occurrence de chaque mot clef défini précédemment (termes d'amorces renseignés par l'utilisateur) et compris dans ce contenu, soit supérieure à un seuil K2.

Dans le présent mode de réalisation le seuil k2=300. A partir de cette valeur les résultats obtenus sont considérés comme étant meilleurs.

Cette étape consiste simplement à vérifier que les quelques mots clés d'amorces sont suffisamment présents dans les deux documents numériques distincts. La valeur de 300 du seuil k2 est un élément permettant de contrôler la qualité des deux documents numériques distincts. Il implique que l'on a trouvé suffisamment de documents numériques initiaux C couvrant le domaine cible (par ex qualité des restaurants, popularité d'un matériel informatique, etc). On remarquera le lien entre cette valeur et celle de la contrainte de taille (100-150ko) évoqué plus haut qui représente à peu prés 20 000 termes par documents numériques distincts(A,B). Les 300 termes représentent 1.5% de ce total. Ainsi en dehors de cas extrêmes (eg. mots amorces, mal orthographiés), les 300 termes sont atteints sans problèmes si la contrainte des 100-150 ko est satisfaite. Cette étape est donc surtout un mode de contrôle. Si cette étape échoue, il convient de refaire une recherche de documents initiaux C. Cette recherche pourra être partielle en ne remplaçant que les 30 % des documents initiaux C dans lesquels les termes d'amorces sont les moins représentés.

[0040] Dans un autre mode de réalisation dit non supervisé, illustré à la figure 4, les documents numériques distincts $((A_i)_{i\epsilon[1\ ;N]})$ et $((B_i)_{i\epsilon[1\ ;N]})$ peuvent être obtenus de manière différente.

En effet, dans cet autre mode de réalisation, l'utilisation de mots clefs d'amorces visant la réalisation d'une recherche pour trouver des documents numériques initiaux C n'est pas nécessaire.

Pour ce faire, le dispositif émet une requête vers la base de données, et de manière alternative vers d'autres bases de données ou encore sur Internet. Cette requête fournit des documents numériques initiaux C se rapportant à un même domaine par exemple la restauration ou encore l'informatique.

Cette recherche vise alors à retrouver des documents numériques initiaux C qui sont par la suite convertis au format de documents textuels contenant ces termes si ces documents numériques initiaux C sont dans des formats différents. Ces documents numériques initiaux C, ici des documents textuels doivent être de même nature et appartenir au même domaine, par exemple, des avis sur des marchands en lignes de produits informatiques.

De plus, il faut que les sources d'informations dont sont issues ces documents numériques initiaux C soient des sources porteuses d'opinions par opposition à des sources uniquement descriptives. Par exemple, les modes d'emplois de smartphones ne peuvent pas convenir car elles sont descriptives et ne contiennent pas d'opinions ou de jugements de valeurs. Par contre un forum, un blog ou encore un site contenant des avis de consommateurs, ou tout autre source réputée (ou éventuellement vérifiée superficiellement) contenir des opinions sur les smartphones peut alors convenir.

Pour cela, ces documents numériques initiaux C font l'objet d'un traitement qui vise à filtrer chacun de ces documents numériques initiaux C en fonction de trois critères dits de « collecte » :

- bipolarité: Ces textes doivent avoir une probabilité « raisonnable » d'être connotés positivement ou négativement. Il n'est pas nécessaire de connaître le sens (positif ou négatif) de cette polarité. Peu importe que certains textes soient neutres mais ils ne doivent pas être majoritaires.

- taille : ces documents numériques initiaux C doivent être en nombre suffisant (10 au minimum) et de tailles proches c'est-à-dire comporter sensiblement le même nombre de termes. Dans certains cas on peut toutefois tolérer une certaine dispersion : par exemple, 10 % des documents numériques initiaux C pourront être de 2 à 3 fois la taille moyenne des

documents numériques initiaux C constituant le résultat de la recherche. Le nombre de documents initiaux C doit être tel que chacun documents (A,B) agrégeant les documents C ait une taille de 100 à 150 ko.

- contextualité : une analyse sémantique, superficielle, de ces documents numériques initiaux C est réalisée afin de vérifier que les éléments constituant chacun des ces documents numériques initiaux C se rapporte à un vocabulaire proche du contexte souhaité. Par exemple l'évaluation des restaurants diffère de l'évaluation de matériel informatique. Cette contrainte de contextualité est naturellement satisfaite lorsque la source d'information initiale est réputée dédiée au contexte choisi. Par exemple, savoir qu'un site web est consacré au partage d'avis d'usagers de Smartphones suffit comme indice de contextualité. Le fait que certains documents dérogent à cette règle n'est pas un problème pourvu qu'ils ne soient pas majoritaires.

Chacun de ces trois critères peut être vérifié superficiellement pour l'ensemble de la source de documents et ce de manière non limitative par un système informatique ou encore un opérateur.

Dès lors seuls les documents numériques initiaux C respectant ces trois critères sont conservés.

[0041] Dans une des étapes mise en œuvre un certain nombre de documents numériques initiaux C collectés sont ensuite concaténés dans les deux documents numériques distincts $((A_i)_{i\epsilon[1\ ;N]})$ et $((B_i)_{i\epsilon[1\ ;N]})$. Ces documents numériques initiaux C gardent leur autonomie car malgré leur répartition et concaténation dans les deux documents numériques distincts $((A_i)_{i\epsilon[1\ ;N]})$ et $((B_i)_{i\epsilon[1\ ;N]})$ il est possible de les isoler individuellement. En effet, ces documents numériques initiaux C sont isolés par des séparateurs textuels dans les documents numériques distincts $((A_i)_{i\epsilon[1\ ;N]})$ et $((B_i)_{i\epsilon[1\ ;N]})$ dans lesquels ils sont concaténés. Mais cet isolement peut être réalisé de différentes manières.

Cette notion d'autonomie est importante car dans le cadre de la mise en œuvre des différents modes de réalisation de l'invention, il faut pouvoir :

– compter le nombre documents numériques initiaux C qui doivent être en nombre identique dans les deux documents numériques distincts $((A_i)_{i\epsilon[1\ ;N]})$ et $((B_i)_{i\epsilon[1\ ;N]})$, et

– noter chacun des documents numériques initiaux C en calculant l'indice I.

Dans le présent cas, un ordinateur est apte à mettre en œuvre les différents modes de réalisation de l'invention.

Ainsi, dans un premier temps, l'ensemble des documents numériques initiaux C sont concaténés aléatoirement à part égale en deux documents numériques distincts $((A_i)_{i\epsilon[1\ ;N]})$ et $((B_i)_{i\epsilon[1\ ;N]})$. Malgré cette concaténation les documents numériques initiaux restent séparables. Cette concaténation n'est aléatoire que dans une première itération. Aux itérations suivantes, la concatenation prendra en compte la valeur de l'indice I qui sera calculé sur chacun des documents initiaux de manière à les classer. Dans ce mode de réalisation dit non supervisé, les deux documents numériques distincts $((A_i)_{i\epsilon[1\ ;N]})$ et $((B_i)_{i\epsilon[1\ ;N]})$ ont une taille de l'ordre de 100 à 150 ko par fichier correspondant à chaque document. Cependant, on notera que statistiquement, les textes connotés négativement sont plus longs que ceux connotés positivement. Comme nous le verrons plus loin, nous mettons à profit ce fait connu de l'état de l'art pour distinguer automatiquement parmi les deux documents A et B, celui qui est connoté négativement.

Ces documents numériques distincts $((A_i)_{i\epsilon[1\ ;N]})$ et $((B_i)_{i\epsilon[1\ ;N]})$ font par la suite l'objet d'un traitement particulier lors d'une étape d'analyse fréquentielle afin d'identifier les termes de leur contenu et leur fréquence d'apparitions, soit leur indice de redondance I1 et I2. Cette étape permettant la détermination des indices de redondance I1 et I2 est similaire à celle mise en œuvre dans le mode de réalisation dit supervisé et qui aboutit à la constitution d'un dictionnaire différentiel

[0042] Par ailleurs, dans l'optique d'améliorer la qualité des documents numériques distincts $((A_i)_{i\epsilon[1\ ;N]})$ et $((B_i)_{i\epsilon[1\ ;N]})$ sur la base desquels est générés le dictionnaire différentiel $((D_i)_{i\epsilon[1\ ;N]})$, il est procédé à une mesure d'indice I sur les documents numériques initiaux C qui ont contribués à l'obtention

dudit dictionnaire différentiel ($(D_i)_{i\in[1;N]}$). L'objectif de cette mesure qui reflète le niveau de connotation (positif/négatif) est de pouvoir mieux classer ensemble les documents initiaux les plus positifs et les séparer des plus négatifs. En effet le document numérique distinct ($(A_i)_{i\in[1;N]}$) ou ($(B_i)_{i\in[1;N]}$) correspond chacun à une concaténation de documents numériques initiaux C connotés négativement ou positivement. Cependant, à ce stade nous ne savons pas lequel des deux documents (A/B) est celui qui est connoté négativement. Le système informatique prend cette décision en se basant sur la taille des fichiers. Cecui qui est le plus long des deux sera celui qui est connoté negativement.

[0043] Dans notre exemple les documents numériques distincts ($(A_i)_{i\in[1;N]}$) et ($(B_i)_{i\in[1;N]}$) qui ont été obtenus sont donc les documents numériques distincts A1 et B1 sur la base desquels le dictionnaire différentiel D1 a été généré.

[0044] A partir notamment de ce dictionnaire différentiel D1, les valeurs d'indice I obtenues pour chacun de ces documents numériques initiaux C permettent de classer ces documents de sorte à obtenir des documents numériques distincts A2 et B2. A2 et B2 correspondant respectivement par exemple à de nouvelles catégories : positive et négative. A chacune de ces itérations on applique le principe précédent : classement à part égale des documents initiaux dans deux fichiers Ai, Bi, le plus long étant désigné comme celui connoté négativement.

[0045] Rappelons qu'une valeur de l'indice I, caractérisant le contenu d'un document numérique, inférieure à 0,5 est un indice de connotation négative et ce d'autant plus que cette valeur de I se rapproche de 0. Lors que cette valeur de I se rapproche de 1, I est alors un indice de connotation positive.

[0046] Le dictionnaire D1 permet donc déjà en termes de probabilité de séparer les documents initiaux C connotée négativement de ceux connotés positivement. En effet, même si dans cette phase initiale, la répartition des documents initiaux C a été aléatoire, les termes positifs (resp negatifs) seront isolés et séparés par le processus de constitution du dictionnaire. Cette capacité de séparation va s'améliorer au fur et à mesure des étapes

suivantes : générations des nouveaux dictionnaires (D2, D3, ..) à partir des reclassements successifs des documents initiaux C.

[0047] Par la suite selon les étapes de génération d'un dictionnaire différentiel $((D_i)_{i\in[1;N]})$ à partir de documents numériques distincts $((A_i)_{i\in[1;N]})$ et $((B_i)_{i\in[1;N]})$ ont obtient alors sur la base des documents numériques distincts A2 et B2 optimisés un dictionnaire différentiel D2.

[0048] Ces étapes visant à l'amélioration de la qualité des documents numériques distincts $((A_i)_{i\in[1;N]})$ et $((B_i)_{i\in[1;N]})$ sont alors réitérées selon un processus de cycles sur la base des nouveaux documents numériques distincts $((A_i)_{i\in[1;N]})$ et $((B_i)_{i\in[1;N]})$ et du dictionnaire différentiel $((D_i)_{i\in[1;N]})$ obtenus jusqu'à ce que la différence de contenu entre deux dictionnaires différentiels successifs soit inférieure à un certain niveau k4.

[0049] On comprend qu'il s'agit ici d'évaluer la différence entre deux dictionnaires successifs. L'algorithme qui produit les dictionnaires converge et au fur et à mesure des itérations, il y a de moins en moins de différences entre deux dictionnaires successifs. Seuls les dictionnaires suffisamment différents sont utiles dans la phase qui suit. Cette différence k4 est exprimée en % de termes différents d'une étape sur l'autre. Dés qu'il y a moins de 5% de mots différents, le processus de création des dictionnaires est stoppé (dictionnaire considéré comme stable).

[0050] Ainsi, ce processus d'amélioration de la qualité des documents numériques distincts $((A_i)_{i\in[1;N]})$ et $((B_i)_{i\in[1;N]})$ s'arrête par exemple si k4=5, c'est-à-dire si moins de 5% des critères constituant un dictionnaire différentiel $((D_i)_{i\in[1;N]})$ ont changé par rapport au dictionnaire différentiel précédent.

[0051] De manière générale, afin d'éviter une oscillation infinie de ce processus le nombre d'itérations est limité à 20 cycles.

[0052] Les différents dictionnaires différentiels $((Di)_{i\in[1;N]})$ obtenus durant ce processus d'amélioration sont alors combinés de sorte à générer un dictionnaire synthétique S.

Cette étape de génération du dictionnaire synthétique S est réalisée selon les sous-étapes suivantes :

- pour chacun des termes contenus dans les N dictionnaires différentiels $((Di)_{i\in[1;N]})$, une somme SG du nombre d'apparitions de chaque terme dans ces N dictionnaires différentiels $((D_i)_{i\in[1;N]})$ est réalisée ;

- pour chacun des termes contenus dans les N dictionnaires différentiels $((Di)_{i\in[1;N]})$, une somme SJ de la valeur de HK en valeur absolue de ce terme dans les N dictionnaires différentiels $((Di)_{i\in[1;N]})$ est réalisée. Le terme pour lequel cette somme SJ est la plus élevée est le terme de fréquence maximale FM possible. On remarquera que la fréquence HK dont il est question ici est une fréquence différentielle. La somme SJ est donc une somme de fréquences différentielles.

Ainsi, un terme donné est ajouté dans le dictionnaire synthétique S, s'il est contenu dans au moins un des N dictionnaires différentiel et si la valeur de la somme SG du nombre d'apparitions de chaque terme dans ces N dictionnaires différentiels $((D_i)_{i\in[1;N]})$, est supérieure à une limite K8 x N.

La valeur K8 =1/4=0,25 car dans cet exemple on ne prend en compte un terme que s'il est présent au moins dans le quart des dictionnaires différentiels de cet ensemble : indication de sa représentativité. Cette valeur peut être modifiée en fonction de la pertinence et de l'optimisation que l'on veut obtenir pour ce dictionnaire synthétique S.

De manière alternative, cette sous-étape de sélection peut être réalisée autrement en prenant en compte pour un terme donnée les conditions suivantes :

- la somme SG est supérieure au produit K8 par le nombre de dictionnaires différentiels compris dans cet ensemble ;

- la somme SJ est supérieure au produit K9 par le max fréquence possible est la fréquence cumulée la plus importante(FM).

FM est la fréquence cumulée maximale constatée, c'est-à-dire celle du terme ayant la fréquence cumulée la plus importante.

La méthode de décision élaboré permet d'abord de collecter les mots les plus présents dans les N dictionnaires (ie cumul d'apparition) et on s'assure que ce sous-ensemble de termes connotés est aussi suffisamment présent dans le vocabulaire en général, c'est-à-dire en fréquence.

De manière générale les valeurs de K8 et K9 correspondent à 0,25 (soit 1/4). Cette valeur peut être modifiée en fonction de la pertinence et de l'optimisation que l'on veut obtenir pour ce dictionnaire synthétique S.

[0053] Ainsi, ce procédé permet d'obtenir un dictionnaire différentiel $((D_i)_{i\epsilon[1;N]})$ et/ ou synthétique S par domaine d'application, dans lequel chaque critère qui la compose comporte un terme avec un indicateur de valence émotionnelle : par exemple le terme « bien» est connoté positivement, donc indicateur hk=+1, le terme « mal » est connoté négativement, donc indicateur hk=-1. Ce dictionnaire différentiel $((D_i)_{i\epsilon[1;N]})$ et/ ou synthétique S sera utilisé comme référence pour mesurer un document numérique M correspondant par exemple à un document textuel inconnu.

En appliquant, à ce document numérique M les critères du dictionnaire différentiel $((D_i)_{i\epsilon[1;N]})$ et/ ou synthétique S on mesure alors un indice I représentatif de la connotation globale du document textuel inconnu.

On notera que l'indice I révélant par exemple la valence émotionnelle ne se limite pas à la simple bipolarité positif/négatif, d'autres sentiments comme : la peur, la confiance, peuvent aussi être mesurés.

[0054] Les dictionnaires différentiels $((D_i)_{i\epsilon[1;N]})$ et/ ou synthétiques S obtenus sont archivés dans la base de données en fonction du domaine auquel ils se rapportent.

[0055] Ainsi qu'on l'aura compris, le présent mode de réalisation trouve une application particulière dans le cadre de la mesure de l'indice I d'un document numérique M révélant par exemple le sentiment qui peut émaner d'un tel document. Un tel outil de mesure d'indice I peut être mis en œuvre

dans bon nombre de services sur Internet (e-réputation, mesure de satisfaction ou de l'expérience client, …).

[0056] Cet outil peut se décliner dans une multitude de contexte différents - vente d'ordinateurs, vente de vêtements - et peut être soumis sans aucune difficulté aux variations linguistiques - changement de langue, évolutions des langues ou usages linguistiques des usagers : langage sms -.

[0057] La présente solution permet tout en gardant de bonnes performances de la mesure de cet indice I de :

- déterminer de manière optimale une mesure fine et précise de la pertinence d'un document numérique M ;

- accélérer la production de dictionnaires différentiels $((Di)_{i\in[1;N]})$ et/ ou synthétiques S adaptés à différent domaine ;

– limiter les traitements manuels qui peuvent être réalisés par des opérateurs, dès lors d'automatiser ce type de processus ;

– limiter les compétences nécessaires - techniques, linguistiques - à la production des dictionnaires différentiels $((Di)_{i\in[1;N]})$ et/ ou synthétiques S ;

- ne pas requérir l'utilisation de dictionnaires formels (par exemple académique)

Revendications

1. Procédé de mesure d'un indice (I) caractérisant le contenu d'un document numérique (M) comportant des éléments descriptifs à partir d'un dictionnaire différentiel $((D_i)_{i\in[1;N]})$ comprenant deux ensembles U_1 et U_2 de critères, lesdits critères se rapportant à des éléments ayant un format similaire à celui des éléments descriptifs de (M) et étant chacun associé à un coefficient de pertinence (HK), ledit procédé comportant les étapes suivantes :

- application des critères desdits ensembles U_1 et U_2 aux éléments descriptifs du contenu dudit document numérique (M) de sorte à déterminer les nombres N1 et N2 d'éléments descriptifs dudit contenu correspondant aux critères desdits ensembles respectivement U_1 et U_2, et

- calcul de l'indice (I) compris dans un intervalle de valeurs [0,1] par l'exécution de l'opération : N1/(N1 + N2).

2. Procédé selon la revendication précédente, caractérisé en ce que ledit ensemble (U1) comporte des critères de même catégorie $(((C_i)_{i\in[1;N]})+HK)$ avec HK>0 et ledit ensemble (U2) comporte des critères de même catégorie $(((C_j)_{j\in[1;N]}) + HK)$ avec HK<0.

3. Procédé selon l'une des revendications précédentes caractérisé en ce que ledit dictionnaire différentiel $((D_i)_{i\in[1;N]})$ est généré à partir des étapes suivantes :

- comparaison entre deux documents numériques sources (S1,S2), comprenant chacun une liste d'éléments associés avec respectivement un indice I_1 et un indice I_2 de redondance ;

- conception de critères.

4. Procédé selon la revendication précédente caractérisé en ce que ladite étape de comparaison comporte l'une des sous-étapes suivantes :

- identification d'un élément compris dans un seul des documents numériques sources (S1,S2), ou
- identification d'un élément compris dans les deux documents numériques sources (S1,S2), et que son indice (I_1, I_2) dans l'un de ces deux documents soit x fois supérieure à son indice (I_1, I_2) dans l'autre de ces deux documents.

5. Procédé selon la revendication précédente caractérisé en ce que la génération du dictionnaire différentiel $((D_i)_{i\epsilon[1;N]})$ comporte une étape de conception de critères par l'association à l'élément identifié lors de l'étape de comparaison d'un coefficient de pertinence HK correspondant à la différence potentiellement négative entre les indices I1 et I2.

6.Procédé selon l'une des revendications 4 ou 5, caractérisé en ce que les deux documents numériques sources (S1,S2) sont générés lors des étapes suivantes :

- concaténation de documents numériques initiaux (C) en deux documents numériques distincts $((A_i)_{i\epsilon[1;N]})$ et $((B_i)_{i\epsilon[1;N]})$, et
- classement des éléments contenus dans chacun de ces documents numériques distincts $((A_i)_{i\epsilon[1;N]})$ et $((B_i)_{i\epsilon[1;N]})$ en fonction de leurs indices de redondances (I_1,I_2).

7. Procédé selon la revendication précédente, caractérisé en ce que les documents numériques initiaux (C) sont extraits de base de données à partir de requêtes comportant des termes d'amorces dont les caractéristiques sémantiques sont opposables..

8. Procédé selon l'une des revendications précédentes 1 à 6, caractérisé en ce que les documents numériques initiaux (C) sont extraits de base de données sans l'utilisation de requêtes comportant des termes d'amorces mais en fonction des critères dits de collectes desdits documents numériques initiaux (C).

9. Procédé selon les revendications 6 à 8, caractérisé en ce que le dictionnaire $((D_i)_{i\epsilon[1;N]})$ généré sur la base des documents numériques distincts $((A_i)_{i\epsilon[1;N]})$ et $((B_i)_{i\epsilon[1;N]})$ est appliqué aux documents numériques initiaux C de sorte à les classer en fonction de leur indice (I) et à générer de nouveaux documents numériques distincts $((A_i)_{i\epsilon[1;N]})$ et $((B_i)_{i\epsilon[1;N]})$.

10. Procédé selon la revendication précédente, caractérisé en ce qu'il comprend la génération d'un dictionnaire $((D_i)_{i\epsilon[1;N]})$ à partir des nouveaux documents numériques distincts $((A_i)_{i\epsilon[1;N]})$ et $((B_i)_{i\epsilon[1;N]})$.

11. Procédé selon la revendication précédente, caractérisé en ce qu'un dictionnaire synthétique (S) est réalisé à partir de plusieurs dictionnaires $((D_i)_{i\epsilon[1;N]})$.

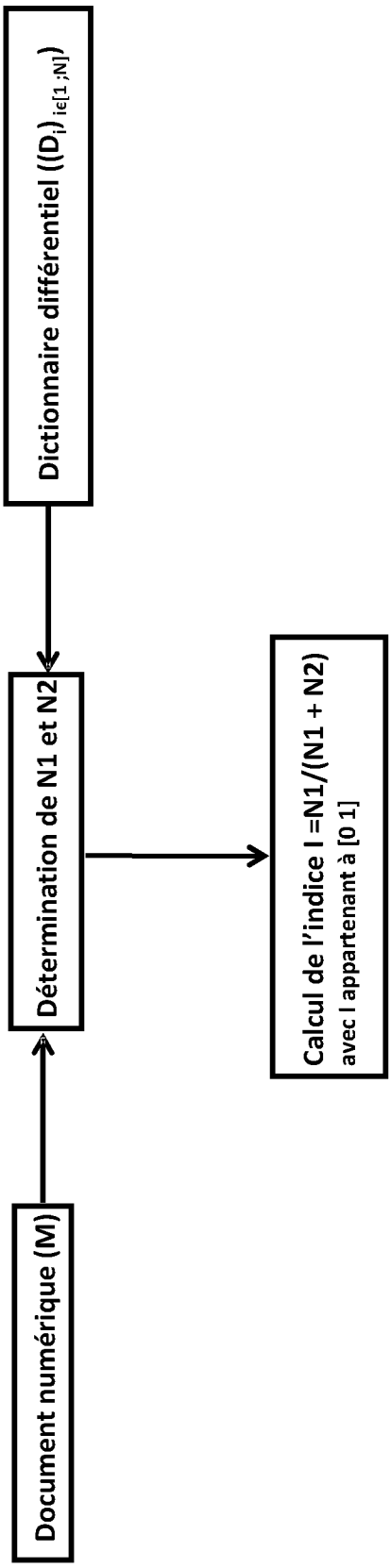
Dictionnaire différentiel $((D_j)_{i\in[1;N]})$
Détermination de N1 et N2
Calcul de l'indice I =N1/(N1 + N2)
avec I appartenant à [0 1]
Document numérique (M)

FIGURE 1

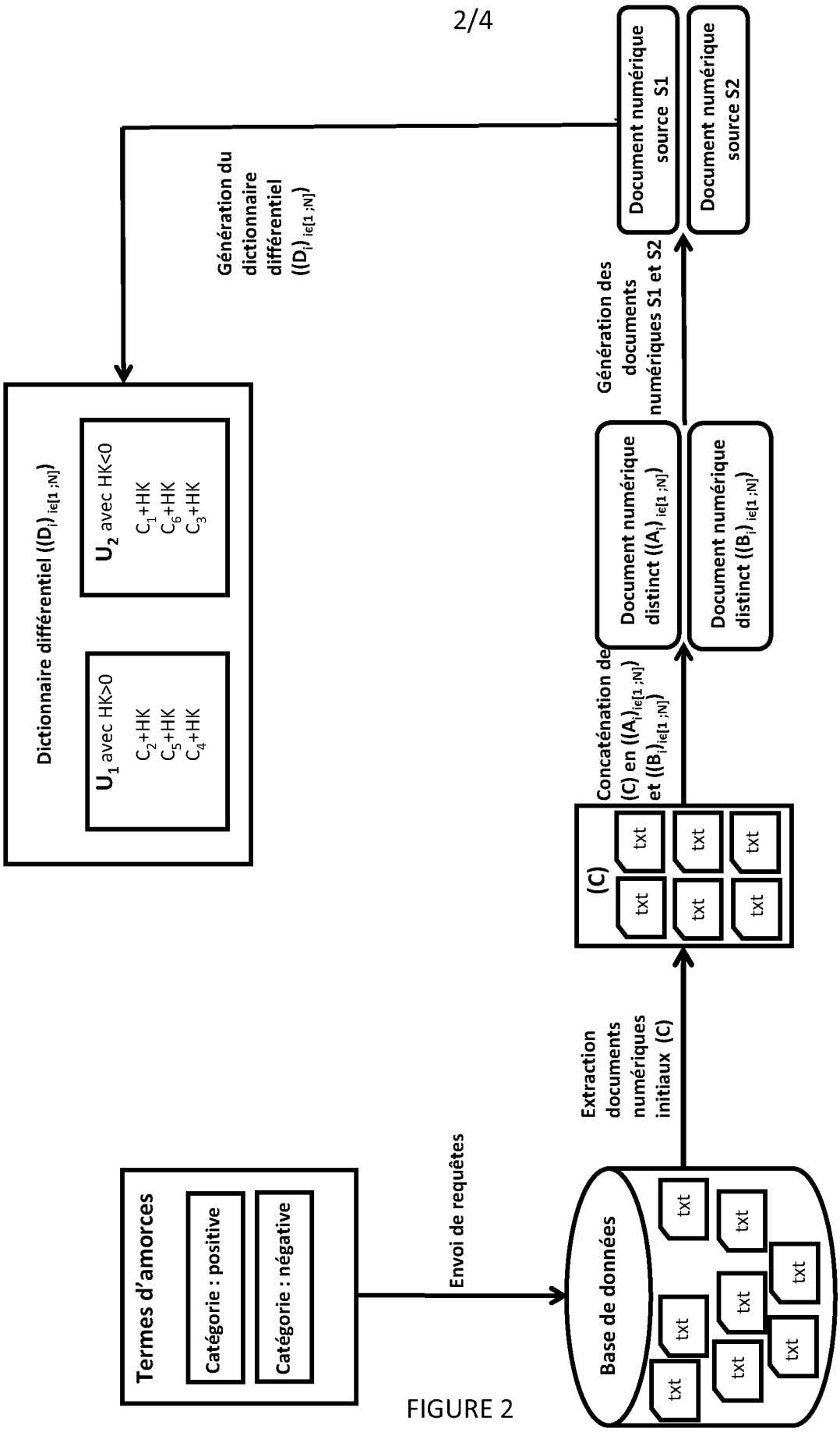
Termes d'amorces
Catégorie : positive
Catégorie : négative
Envoi de requêtes
Base de données
txt
Extraction documents numériques initiaux (C)
(C)
Concaténation de (C) en $((A_i)_{i\in[1;N]})$ et $((B_i)_{i\in[1;N]})$
Document numérique distinct $((A_i)_{i\in[1;N]})$
Document numérique distinct $((B_i)_{i\in[1;N]})$
Génération des documents numériques S1 et S2
Document numérique source S1
Document numérique source S2
Génération du dictionnaire différentiel $((D_i)_{i\in[1;N]})$
Dictionnaire différentiel $((D_i)_{i\in[1;N]})$
U_1 avec HK>0
C_2+HK
C_5+HK
C_4+HK
U_2 avec HK<0
C_1+HK
C_6+HK
C_3+HK

FIGURE 2

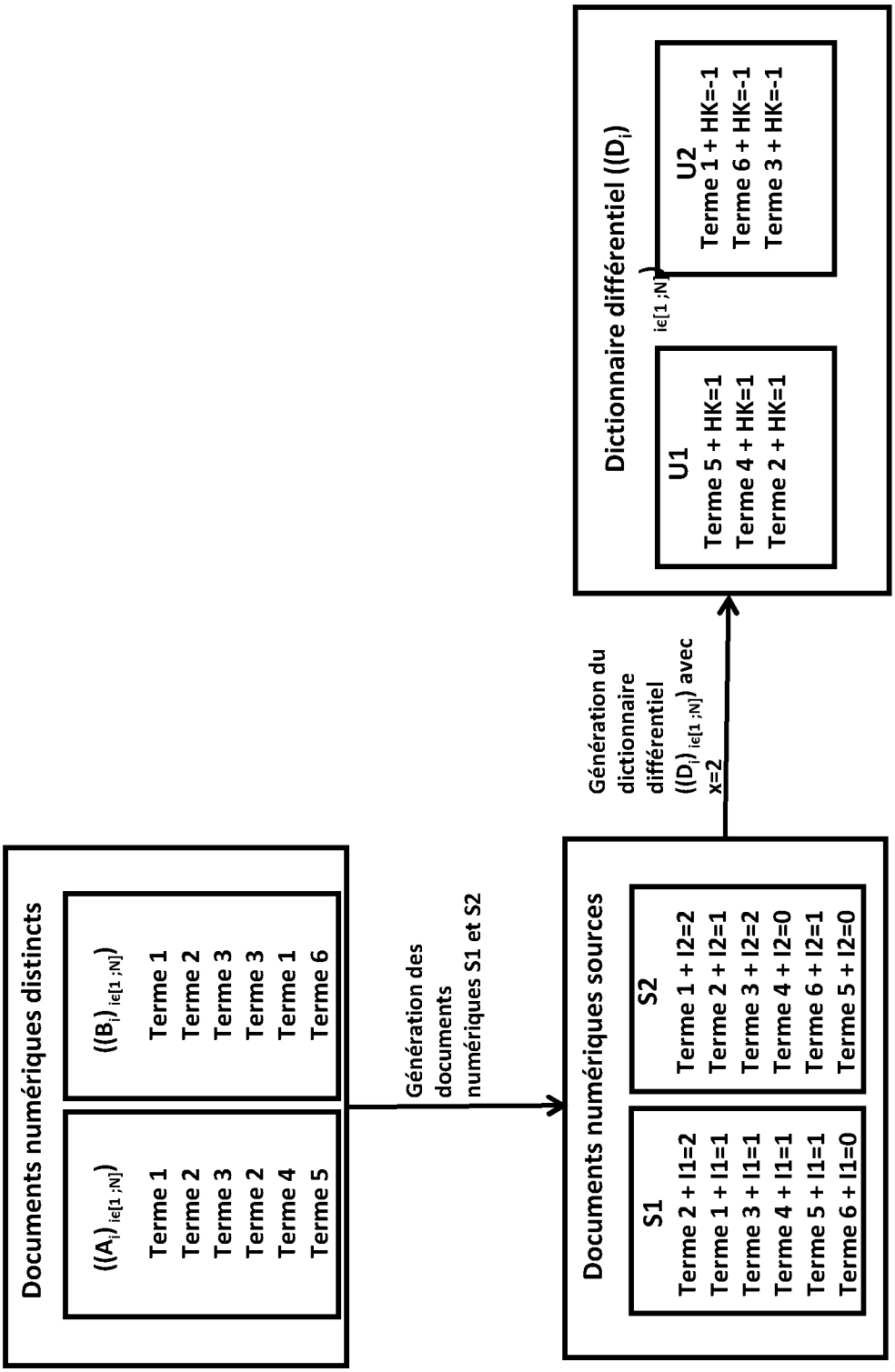
Documents numériques distincts
((A_i)_{i∈[1 ;N]})
Terme 1
Terme 2
Terme 3
Terme 2
Terme 4
Terme 5
((B_i)_{i∈[1 ;N]})
Terme 1
Terme 2
Terme 3
Terme 3
Terme 1
Terme 6
Génération des documents numériques S1 et S2
Documents numériques sources
S1
Terme 2 + l1=2
Terme 1 + l1=1
Terme 3 + l1=1
Terme 4 + l1=1
Terme 5 + l1=1
Terme 6 + l1=0
S2
Terme 1 + l2=2
Terme 2 + l2=1
Terme 3 + l2=2
Terme 4 + l2=0
Terme 6 + l2=1
Terme 5 + l2=0
Génération du dictionnaire différentiel ((D_i)_{i∈[1 ;N]}) avec x=2
Dictionnaire différentiel ((D_i)_{i∈[1 ;N]})
U1
Terme 5 + HK=1
Terme 4 + HK=1
Terme 2 + HK=1
U2
Terme 1 + HK=-1
Terme 6 + HK=-1
Terme 3 + HK=-1

FIGURE 3

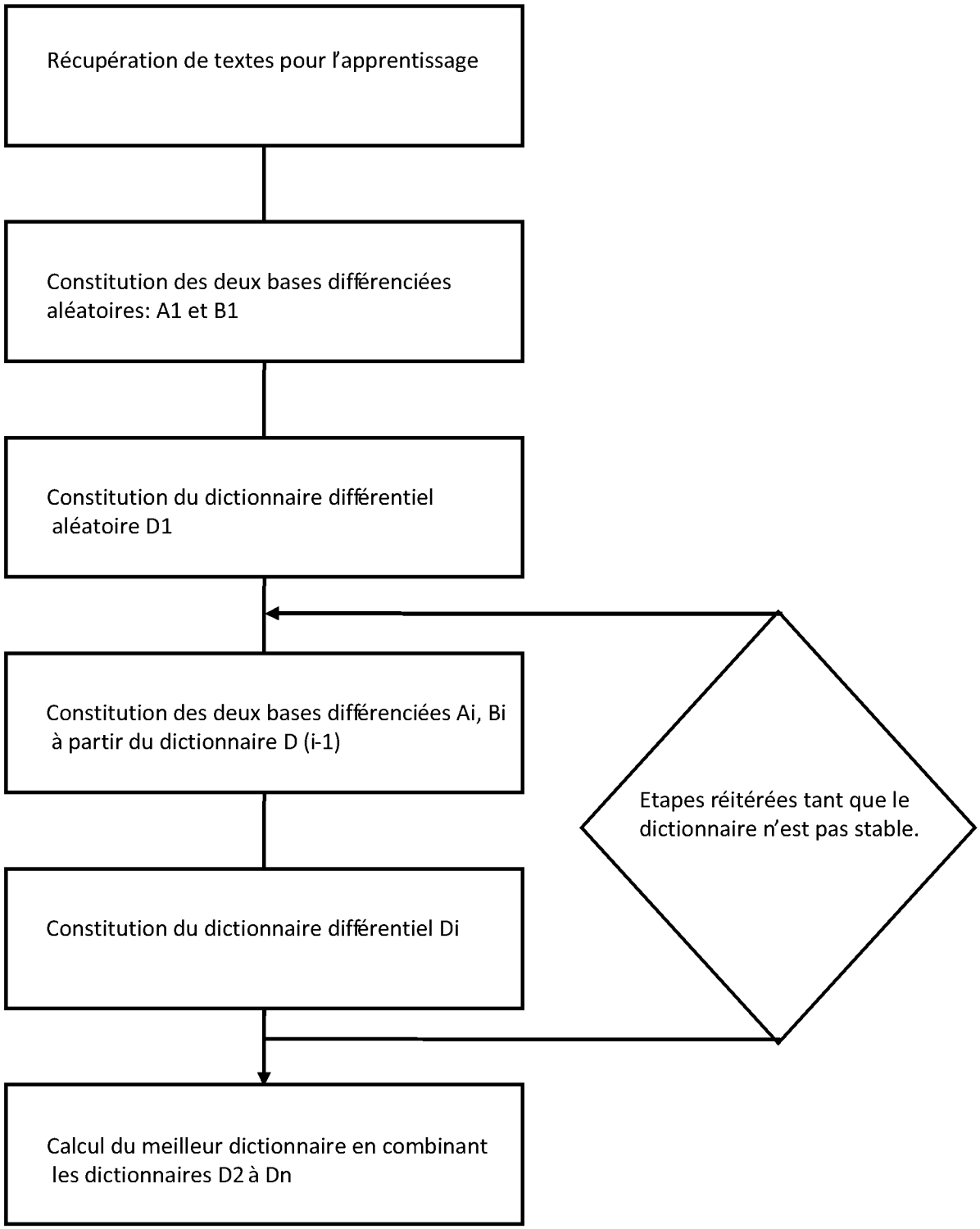
Récupération de textes pour l'apprentissage
Constitution des deux bases différenciées aléatoires: A1 et B1
Constitution du dictionnaire différentiel aléatoire D1
Constitution des deux bases différenciées Ai, Bi à partir du dictionnaire D (i-1)
Constitution du dictionnaire différentiel Di
Etapes réitérées tant que le dictionnaire n'est pas stable.
Calcul du meilleur dictionnaire en combinant les dictionnaires D2 à Dn

FIGURE 4