(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
6 août 2015 (06.08.2015)

WIPO | PCT

(10) Numéro de publication internationale
WO 2015/114014 A1

(51) Classification internationale des brevets :
G06F 17/27 (2006.01) *G06F 17/30* (2006.01)

(21) Numéro de la demande internationale :
PCT/EP2015/051722

(22) Date de dépôt international :
28 janvier 2015 (28.01.2015)

(25) Langue de dépôt : français

(26) Langue de publication : français

(30) Données relatives à la priorité :
1400201 28 janvier 2014 (28.01.2014) FR

(71) Déposant : **DEADIA** [FR/FR]; 10 rue de la Renaissance, 92160 Antony (FR).

(72) Inventeur : **MALLE, Jean-Pierre**; 16, rue Bizet, 91240 Saint Michel S/Orge (FR).

(74) Mandataire : **REGIMBEAU**; 20, rue de Chazelles, 75847 Paris Cedex 17 (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, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, 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, SA, 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, ST, 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).

Déclarations en vertu de la règle 4.17 :

— *relative à la qualité d'inventeur (règle 4.17.iv))*

Publiée :

— *avec rapport de recherche internationale (Art. 21(3))*

(54) **Title** : METHOD FOR SEMANTIC ANALYSIS OF A TEXT

(54) **Titre** : PROCÉDÉ D'ANALYSE SÉMANTIQUE D'UN TEXTE

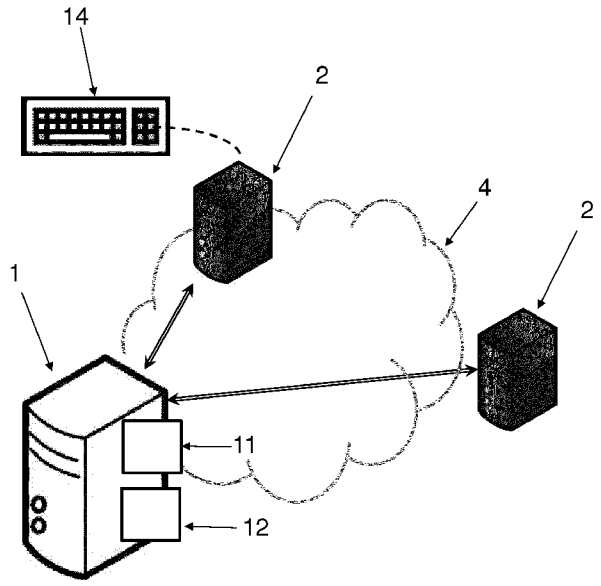

FIG. 1

(57) **Abstract** : The present invention relates to the field of computer-based semantic understanding. Specifically, it relates to a method for semantic analysis of a natural-language text by data-processing means with a view to the classification thereof.

(57) **Abrégé** : La présente invention concerne le domaine de la compréhension sémantique par ordinateur. Plus précisément elle concerne un procédé d'analyse sémantique d'un texte en langage naturel par des moyens de traitement de données, en vue de sa classification.

WO 2015/114014 A1

Procédé d'analyse sémantique d'un texte

DOMAINE TECHNIQUE GENERAL

La présente invention concerne le domaine de la compréhension sémantique par ordinateur.

Plus précisément elle concerne un procédé d'analyse sémantique d'un texte en langage naturel par des moyens de traitement de données, en vue de sa classification.

ETAT DE L'ART

L'analyse sémantique d'un texte en langage naturel vise à en établir la signification en utilisant le sens des mots qui le constituent, suite à une analyse lexicale qui permet de décomposer ce texte à l'aide d'un lexique ou d'une grammaire. L'humain le réalise inconsciemment pour comprendre les textes qu'il lit, et des développements récents visent à conférer des capacités semblables aux machines.

On connait pour le moment des algorithmes d'analyse sémantique automatisée conçus pour qu'un ordinateur puisse classer un texte dans plusieurs catégories prédéterminées, par exemple des thèmes généraux tels que « nature », « économie », « littérature », etc.

Toutefois, cette classification s'avère très limitée et peu évolutive. Dans la mesure où le choix des diverses catégories possible est souvent arbitraire, des textes situés à la frontière de deux catégories peuvent poser problème aux algorithmes. De plus, classifier plus finement dégrade fortement les performances des algorithmes et entraîne des erreurs d'appréciation, causées notamment par les ambiguïtés dues à certains homonymes et certaines tournures (par exemple une double négation).

De façon générale, donner par un traitement informatique un sens « absolu » à un texte est une opération très complexe et souvent contestable. Par exemple, déterminer si un texte prend position « pour »

ou « contre » une opinion est aujourd'hui hors de portée de l'analyse sémantique informatisée.

Il serait souhaitable de disposer d'un procédé amélioré d'analyse sémantique d'un texte par un ordinateur en vue de sa classification qui soit significativement plus performant et plus fiable que tout ce qui fait actuellement, et qui ne soit pas limité par des modèles sémantiques préétablis.

PRESENTATION DE L'INVENTION

La présente invention propose un procédé d'analyse sémantique d'un texte en langage naturel reçu par un équipement depuis des moyens de saisie, le procédé étant caractérisé en ce qu'il comprend la mise en œuvre par des moyens de traitement de données de l'équipement d'étapes de :

(a) Découpage syntaxique d'au moins une partie du texte en une pluralité de mots ;

(b) Filtrage des mots de ladite partie de texte par rapport à une pluralité de liste de mots de référence stockées sur des moyens de stockage de données de l'équipement chacune étant associée à une thématique, de sorte à identifier :

- L'ensemble des mots de ladite partie du texte associés à au moins une thématique,
- L'ensemble des thématiques de ladite partie du texte ;

(c) Construction d'une pluralité de sous-ensembles de l'ensemble des mots de ladite partie du texte associés à au moins une thématique ;

(d) Pour chacun desdits sous-ensembles et pour chaque thématique identifiée, calcul :

- d'un coefficient de couverture de la thématique et/ou d'un coefficient de pertinence de la thématique en fonction d'occurrences dans ladite partie du texte de mots de référence associés à la thématique ;

- d'au moins un coefficient d'orientation de la thématique à partir des mots de ladite partie du texte ne faisant pas partie du sous-ensemble ;

(e) Pour chacun desdits sous-ensembles et pour chaque thématique identifiée, calcul d'un coefficient sémantique représentatif d'un degré de sens porté par le sous-groupe en fonction desdits coefficients de couverture, pertinence et/ou orientation de la thématique.

(f) Sélection en fonction des coefficients sémantiques d'au moins un couple sous-ensemble/thématique.

(g) Classification du texte en fonction dudit au moins un couple sous-ensemble/thématique sélectionné.

Selon d'autres caractéristiques avantageuses et non limitatives de l'invention :

- un coefficient de couverture d'une thématique est calculé à l'étape (d) comme le nombre *N* de mots de référence associés à la thématique compris dans ledit sous-ensemble ;
- un coefficient de pertinence d'une thématique est calculé à l'étape (d) par la formule $N * (1 + \ln(R))$, où *N* est le nombre de mots de référence associés à la thématique compris dans le sous-ensemble et *R* le nombre total d'occurrences dans ladite partie du texte de mots de référence associés à la thématique ;
- deux coefficients d'orientation de la thématique sont calculés à l'étape (c), dont un coefficient de certitude de la thématique et un coefficient de nuance de la thématique ;
- un coefficient de certitude d'une thématique est calculé à l'étape (d) comme valant :
  - 1 si les mots ne faisant pas partie du sous-ensemble sont représentatifs d'une proximité affirmative avec la thématique ;
  - -1 si les mots ne faisant pas partie du sous-ensemble sont représentatifs d'une proximité négative avec la thématique ;

- 0 si les mots ne faisant pas partie du sous-ensemble sont représentatifs d'une proximité incertaine avec la thématique ;

• un coefficient de nuance d'une thématique est un scalaire positif supérieur à 1 lorsque les mots ne faisant pas partie du sous-ensemble sont représentatifs d'une amplification de la thématique, et un scalaire positif inférieur à 1 lorsque les mots ne faisant pas partie du sous-ensemble sont représentatifs d'une atténuation de la thématique ;

• le procédé comprend une étape (a0) préalable de découpage du texte en une pluralité de proposition, chacune étant une partie du texte pour laquelle les étapes (a) à (d) du procédé selon répétées de sorte à obtenir pour chaque proposition un ensemble de coefficients de couverture, de pertinence, et/ou d'orientation associés à la proposition, le procédé comprenant préalablement à l'étape (e) une étape (e0) de calcul pour chacun desdits sous-ensemble et pour chaque thématique identifiée pour au moins une proposition du texte d'un coefficient global de couverture de la thématique et/ou d'un coefficient global de pertinence de la thématique, et d'au moins un coefficient global d'orientation de la thématique en fonction de l'ensemble desdits coefficients associés une proposition ;

• un coefficient global de couverture d'une thématique est calculé à l'étape (e0) comme la somme des coefficients de couverture de la thématique associée à une proposition moins le nombre de mots de référence de la thématique présents dans au moins deux propositions ;

• un coefficient global de pertinence d'une thématique est calculé à l'étape (e0) comme la somme des coefficients de pertinence de la thématique associée à une proposition ;

• un coefficient global d'orientation d'une thématique est calculé à l'étape (e0) comme la moyenne des coefficients d'orientation de la thématique associés à une proposition pondérés par les coefficients de couverture de la thématique associés ;

• l'étape (e0) comprend pour chacun desdits sous-ensembles et pour chaque thématique le calcul d'un coefficient global de divergence de la thématique correspondant à l'écart-type de la distribution des produits des

coefficients d'orientation par les coefficients de couverture associés à chaque proposition ;

- un coefficient sémantique d'un sous-ensemble *A* pour une thématique *T* est calculé à l'étape (e) par la formule M(A,T) = coefficient de pertinence(A,T) * coefficient d'orientation(A,T) * $\sqrt{}$ [1 + coefficient de divergence(A,T)^2] ;
- les couples sous-ensemble/thématique sélectionnés à l'étape (f) sont ceux tels que pour toute partition du sous-ensemble en une pluralité de parties dudit sous-ensemble, le coefficient sémantique du sous-ensemble pour la thématique est supérieur à la somme des coefficients sémantiques des sous-parties du sous-ensemble pour la thématique ;
- des groupes de couples sous-ensemble/thématique de référence sont stockés sur les moyens de stockage de données, l'étape (g) comprenant la détermination du ou des groupes comprenant au moins un couple sous-ensemble/thématique sélectionné à l'étape (f) ;
- l'étape (g) comprend la création d'un nouveau groupe si aucun groupe de couples sous-ensemble/thématique de référence ne contient au moins un couple sous-ensemble/thématique sélectionné pour le texte ;
- chaque couple sous-ensemble/thématique de référence est associé à un score stocké sur les moyens de stockage de données, le score d'un couple sous-ensemble/thématique de référence diminuant avec le temps mais augmentant à chaque fois que ce couple sous-ensemble/thématique est sélectionné pour un texte ;
- le procédé comprend une étape (h) de suppression d'un couple sous-ensemble/thématique de référence d'un groupe si le score dudit couple passe en dessous d'un premier seuil, ou de modification sur les moyens de stockage de données (12) de ladite pluralité de listes associées aux thématiques si le score dudit couple passe au-dessus d'un deuxième seuil ;
- l'étape (g) comprend pour chaque groupe de couples sous-ensemble/thématique de référence le calcul d'un coefficient de dilution représentant le nombre d'occurrences dans ladite partie du texte de mots

de référence associés à des thématiques des couples sous-ensemble/thématique de référence présents dans le texte rapporté au nombre total de mots de référence associés auxdites thématiques ;

- tous les sous-ensembles de l'ensemble des mots de ladite partie du texte associés à au moins une thématique sont construits à l'étape (c).

Selon un deuxième aspect, l'invention concerne un équipement comprenant des moyens de traitement de données configurées pour mettre en œuvre suite à la réception d'un texte en langage naturel un procédé selon le premier aspect de l'invention d'analyse sémantique du texte.

BREVE DESCRIPTION DES FIGURES

D'autres caractéristiques et avantages de la présente invention apparaîtront à la lecture de la description qui va suivre d'un mode de réalisation préférentiel. Cette description sera donnée en référence aux dessins annexés dans lesquels :

- la figure 1 est un schéma d'une architecture réseau dans laquelle s'inscrit l'invention ;
- la figure 2 est un diagramme représentant schématiquement les étapes du procédé d'analyse sémantique selon l'invention.

DESCRIPTION DETAILLEE D'UN MODE DE REALISATION PREFERE

Architecture

En référence à la **figure 1**, le présent procédé est mis en œuvre par des moyens de traitement de données 11 (qui consistent typiquement en un ou plusieurs processeurs) d'un équipement 1. Ce dernier peut être par exemple un ou plusieurs serveurs connectés à un réseau 4, typiquement internet, via lequel il est relié à des clients 2 (par exemple des PC personnels).

L'équipement 1 comprend en outre des moyens de stockage de données 12 (typiquement un ou plusieurs disques durs).

La notion de texte

Un texte est ici n'importe quel message en langage naturel et porteur de sens. Le texte est reçu sous forme électronique, c'est-à-dire en un format directement traitable par les moyens de traitement 11, par exemple XML (eXtensible Markup Language). On comprendra que par reçu « depuis des moyens de saisie 14 », on entend une grande variété d'origines. De façon générale, le terme moyens de saisie désigne tout moyens, hardware et/ou software, permettant de récupérer le texte et de l'envoyer aux moyens de traitement de données 11 sous un format lisible. Le texte peut être directement tapé par un utilisateur, et les moyens de saisie 14 désignent par exemple un clavier et un logiciel de traitement de texte. Alternativement, le texte peut être un texte papier scanné et reconnu par OCR (reconnaissance optique de caractères), et les moyens de saisie 14 désignent alors un scanner et un logiciel de traitement des données numérisées, ou encore le texte peut être dicté et les moyens de saisie 14 désignent alors un microphone et un logiciel de reconnaissance vocale. Enfin, le texte peut être reçu par exemple depuis un serveur du réseau internet, éventuellement directement sous un format lisible. Le présent procédé n'est limité à aucun type de texte. Dans une structure connectée du type de la figure 1, les moyens de saisie sont typiquement ceux d'un client 2 ou un autre serveur 1.

Le texte est structuré en sections. Les sections peuvent être séparées par des paragraphes ou êtres simplement enchainées. Les sections se distinguent les unes des autres par le fait que les concepts exposés sont sensiblement différents. La détection des sections non marquées par l'auteur est une opération complexe.

Une section est composée de phrases séparées par une ponctuation (deux points, point, point d'exclamation, point d'interrogation, tiret d'alinea, points de suspension, etc.).

Une phrase est composée de propositions séparées par une ponctuation (virgule, point-virgule).

Une proposition est une suite de mots séparés par des espaces.

Un mot est un ensemble ordonné de lettres et de signes particuliers (accents, tirets, etc.).

Dans certains textes, les ponctuations peuvent ne pas être respectées. Certains textes peuvent contenir des mots abrégés ou des mots éludés.

Dans une première étape (a), dite de « parsing », au moins une partie du texte est découpée syntaxiquement en une pluralité de mots. Avantageusement, cette partie de phrase est une proposition, et le texte est d'abord découpé proposition par proposition dans une étape (a0) avant que chaque proposition soit tour à tour découpée en mots. On connait des algorithmes capables, notamment grâce à des règles de grammaire, d'identifier les propositions. Le découpage par propositions peut se faire suite à un découpage par phrases, lui-même après un découpage par sections. L'identification des mots se fait grâce aux espaces.

Typiquement, un parseur (le moteur mettant en œuvre le parsing) utilisant la ponctuation et la mise en forme comme délimiteur des propositions peut suffire si les ponctuations sont respectées.

Au sein d'une proposition, l'homme du métier utilisera par exemple un parseur mettant en œuvre les règles suivantes :

- remplacement de chaque verbe par ce verbe à l'infinitif et association à ce dernier de trois indices (le mode, le temps, la personne) ;
- remplacement de chaque nom par ce nom au singulier et association à ce dernier de deux indices (le genre, le nombre) ;
- remplacement de chaque adjectif par cet adjectif au masculin singulier et association à ce dernier de deux indices (le genre, le nombre) ;

- conservation des adverbes ;
- suppressions des mots « enjoliveurs » de la langue (à l'aide d'une liste) ;
- déclaration comme nom propre de tout autre terme ;
- inscription de chaque mot, son type et ses indices dans une liste associée à la proposition.

Les présentes règles peuvent être modifiées ou supprimées, d'autres règles peuvent enrichir le parseur.

La notion de catégories et thématiques

Un texte se classe dans une ou plusieurs « catégories » en fonction du sens qu'il porte. Les catégories sont ici des ensembles mouvants.

Comme l'on verra plus loin, les catégories sont définies comme des groupes « d'anneaux » et peuvent être induites par l'apparition d'un texte relevant d'un sens nouveau.

Lorsqu'une catégorie devient trop peuplée il est souhaitable de la segmenter en réduisant le spectre des sens admissibles dans chaque groupe de textes formés par la scission de la catégorie initiale. Chaque groupe de texte devient alors une catégorie. Une catégorie se représente par une liste de thématiques.

Le thème est le sens est attaché à un ensemble de mots (dits mots de référence) entrant dans la composition d'une proposition, présents dans une liste appelée thématique. La thématique est attachée à une ou plusieurs catégories.

Pour chaque thématique, la liste des mots de référence associée est stockée sur les moyens de stockage 12 de l'équipement 1.

Par exemple, une thématique « motorisation » peut comprendre des mots de référence {moteur, piston, cylindre, vilebrequin, arbre, bielle, pédale, puissance, etc.}, et une thématique « géométrie » peut comprendre les mots de référence {droite, angle, degré, étoile, rectangle, sphère, cylindre, pyramide, etc.}. On voit notamment que le mot « cylindre »

présente plusieurs sens et est ainsi lié aux deux thématiques bien qu'elles soient éloignées.

Dans la suite de la présente description, on prendra l'exemple d'une proposition formulée comme suit : « le moteur comprend trois pistons reliés à un vilebrequin par des bielles en étoile formant un angle de 120° deux à deux qui réagit à la moindre pression sur la pédale d'accélération », ou de légères variation de cette proposition.

Dans l'étape (b), au moins une thématique est identifiée parmi la pluralité de thématiques chacune associées à une liste de mots de référence de la thématique stockée.

En particulier, il suffit qu'un mot de référence associé à la thématique soit présent pour que la thématique soit associée. Alternativement, au moins deux (voire plus) mots sont requis.

Dans notre exemple :

- le groupe de mots {moteur, piston, vilebrequin, bielle, pédale} permet de détecter une thématique « motorisation »
- le groupe de mots {angle, 120°, étoile} permet de détecter une thématique « géométrie ».

L'ensemble des mots de la partie du texte analysée associés à au moins une thématique est également identifié. Il s'agit ici de {moteur, piston, vilebrequin, bielle, pédale, angle, 120°, étoile}

Anneaux sémantiques

Soit V un vocabulaire de *Nv* mots (en particulier l'ensemble des mots de référence d'au moins une thématique).

Soit T un sous ensemble de V de *Nt* mots (en particulier l'ensemble des mots de référence présents dans au moins une thématique), $Nt \leq Nv$.

Soit P une proposition de *Np* mots, telle que $Np \leq Nv$.

Soit Q le groupe de *Nq* mots communs à P et à T (il s'agit des mots de la proposition appartenant à au moins une thématique), $Nq \leq Np$.

Soit **P**(P) l'ensemble des parties de P et **P**(Q) l'ensemble des parties de Q.

Par construction, **P**(P) et **P**(Q) sont des anneaux commutatifs unitaires munis de deux opérateurs :

- un opérateur de différence symétrique noté Δ (relativement à deux ensembles A et B, la différence symétrique de A et B est l'ensemble contenant les éléments contenus dans A mais pas dans B, et les éléments contenus dans B et pas dans A) ; et
- un opérateur d'intersection noté &.

P(P) est isomorphe à Z/NpZ et P(Q) est isomorphe à Z/NqZ

$\forall A \in \boldsymbol{P}(P)$, **P**(A) est inclus dans **P**(P) et A est aussi un anneau commutatif unitaire. A contient toutes les combinaisons complètes ou partielles d'un groupe de mots. On appelle A un « anneau sémantique ». A partir de l'ensemble des mots d'une proposition appartenant à une thématique, un anneau sémantique est défini par un sous-ensemble de cet ensemble.

Par exemple, si « ce véhicule est grand et bleu » est une proposition, les anneaux sémantiques de cette proposition sont notés {}, {véhicule}, {grand}, {bleu}, {véhicule, grand}, {véhicule, bleu}, {véhicule, grand, bleu}. Il est important de comprendre que chaque anneau n'est pas la simple liste des mots qui le compose, mais bien l'ensemble des ensembles comprenant $i \in [\![0, K]\!]$ de ces mots (qui sont d'autres anneaux sémantiques). Par exemple, l'anneau défini par véhicule et grand correspond en réalité à l'ensemble { {} ; {véhicule} ; {grand} ; {véhicule, grand} }.

Un anneau est dit centré s'il n'existe pas deux mots qu'il contienne appartenant à deux thématiques différentes (mais il peut contenir des mots n'appartenant à aucun thématique).

Un anneau est dit régulier s'il appartient aussi à **P**(Q), c'est-à-dire que tous les mots qu'il contient appartiennent à l'une des thématiques.

Dans une étape (c), le procédé comprend la construction d'une pluralité de sous-ensembles de l'ensemble des mots de ladite partie du texte associés à au moins une thématique, en d'autres termes les anneaux

sémantiques réguliers, et avantageusement le procédé comprend la construction de la totalité de ces anneaux.

Si l'ensemble des mots associés à au moins une thématique comprend K éléments, alors il y a 2^K anneaux construits.

Matrices sémantiques

Dans l'étape (d), une représentation du « sens » des anneaux sémantique d'une partie du texte (qui comme expliqué est typiquement une proposition) est déterminée par les moyens de traitement de données 11 de l'équipement 1. Cette représentation prend la forme d'une matrice formée de vecteurs attachés aux thématiques et comprenant plusieurs dimensions et stockée dans les moyens de stockage de données 12 de l'équipement. Cette matrice est appelée « matrice sémantique » (ou matrice de sens). Dans l'hypothèse d'un traitement proposition par proposition, une suite de matrices sémantiques est déterminée, et dans une étape (e0) une matrice sémantique globale du texte est déterminée en fonction des matrices sémantiques des anneaux des propositions.

Une matrice sémantique comprend au moins deux dimensions, avantageusement trois, voire quatre : la couverture, la pertinence (au moins une parmi ces deux est requise), la certitude, la nuance (les deux dernières peuvent être regroupées en une seule dimension, l'orientation). La matrice globale d'un texte peut comprendre une cinquième dimension (la divergence).

Coefficient de couverture d'une thématique

Le procédé comprend pour chaque sous-groupe (i.e. anneau sémantique) et chaque thématique identifiée, le calcul dans d'un coefficient de couverture de la thématique et/ou d'un coefficient de pertinence de la thématique (avantageusement les deux), en fonction d'occurrences dans l'anneau de mots de référence associés à la thématique.

Le coefficient de couverture d'une thématique matérialise la proximité entre l'anneau et la thématique, et se représente par un nombre entier, typiquement le nombre *N* de mots de la thématique compris dans l'anneau. Il est possible d'adjoindre des pondérations (par exemple à certains mots « essentiels » de la thématique).

Dans l'exemple précédent, la proximité entre la proposition et la thématique « motorisation » est plus forte que celle avec la thématique « géométrie » (coefficient de cinq contre trois).

Coefficient de pertinence d'une thématique

Le coefficient de pertinence est calculé par les moyens de traitement de données 11 comme le coefficient de couverture mais en prenant en compte le nombre total d'occurrence des mots du thème.

En particulier, si *N* est le nombre de mots de la thématique contenus dans l'anneau, ou chaque mot ne compte qu'une fois (en d'autres termes le coefficient de couverture de la thématique) et *R* est le nombre de mots de la thématique contenus dans l'anneau, ou chaque mot compte autant de fois qu'il apparait dans la proposition (nombre d'occurrence total, qui croit avec la longueur de la proposition), le coefficient de pertinence est par exemple donné par la formule $N * (1 + \ln(R))$, avec ln le logarithme népérien.

Le calcul d'un coefficient de pertinence n'est pas limité à cette formule, et l'homme du métier pourra par exemple utiliser les formules $ch(\frac{R}{N})$ avec ch le cosinus hyperbolique, ou encore $\frac{1}{\pi} * atan(\frac{R}{N})$ avec atan l'arc tangente, selon le nombre et la taille des thématiques existantes. Chacune de ces formules peut être normalisée.

L'utilisation de l'arc tangente amortit l'effet des grandes valeurs de R, alors qu'on contraire le cosinus hyperbolique accentue l'effet des grandes valeurs de R.

Coefficient de certitude d'une thématique

Le procédé comprend également le calcul, toujours pour chaque sous-groupe (i.e. anneau sémantique) et chaque thématique identifiée, d'au moins un coefficient d'orientation de la thématique à partir des mots de ladite partie du texte ne faisant pas partie de l'anneau (en particulier ceux n'appartenant à aucun anneau).

En particulier, deux coefficients d'orientation de la thématique sont calculés à l'étape (d), dont un coefficient de certitude de la thématique et un coefficient de nuance de la thématique.

La certitude est véhiculée par un ensemble de mots dont l'ordre et la nature peut changer radicalement le sens porté par la proposition. Il s'agit typiquement des mots tels que des négations, de la ponctuation, des mots interrogatifs/négatifs, dont une liste peut être stockée sur les moyens de stockage de données 12. La position de ces mots les uns par rapport aux autres (typique de certaines tournures) donne par ailleurs des indices sur la certitude.

Selon ces mots, la proximité peut être affirmative, négative ou incertaine. Dans l'exemple précédent, la proximité est affirmative (faute de mots modifiant la certitude).

Par comparaison, dans une proposition qui serait formulée « aucun moteur ne comprenant aucune bielle ni aucun piston n'équipe ce véhicule à pédale », la motorisation est une anti-thématique, révélée par les mots répétés « aucun(e) », « ni » et « n' ».

La proximité entre ce texte et la thématique « motorisation » est négative.

Par comparaison encore, dans l'exemple : « ce véhicule serait-il équipé d'un moteur à piston et d'un vilebrequin à bielles ? », la proximité entre le texte et la catégorie « motorisation » est interrogative du fait de la tournure interrogative et la présence du point d'interrogation.

La certitude peut ainsi se représenter par trois valeurs :

- 1 pour l'affirmative

- -1 pour la négative
- 0 pour l'incertitude (interrogatif, interronégatif, affirmatif et négatif entremêlés, etc.)

Coefficient de nuance d'une thématique

La nuance est véhiculée par un ensemble de mots dont l'ordre et la nature peut altérer le sens porté par la proposition. Cette altération peut être un renforcement ou un affaiblissement de la proximité avec la thématique, par exemple grâce à des adverbes tels que « certainement », « assurément », « probablement », « éventuellement ». Comme pour la nuance, il est possible de stocker sur les moyens de stockage 12 une liste des mots caractéristiques d'un renforcement ou d'un affaiblissement de la proximité avec une thématique. Les moyens de traitement de données 11 comparent les mots non associés avec la thématique avec cette liste et en déduisent la valeur du coefficient de nuance, qui est en particulier un scalaire positif (supérieur à 1 pour un renforcement et inférieur à 1 pour un affaiblissement)

Dans l'exemple : « Assurément ce moteur comprend bien un vilebrequin et des bielles, » la nuance est un renforcement de la thématique (grâce à « assurément »), et le coefficient est par exemple 1.5.

Dans l'exemple : « Matthieu croit savoir que le moteur contient un vilebrequin et des bielles, » la nuance est un affaiblissement de la thématique (grâce à « croire »), et le coefficient est par exemple 0.75.

Il est à noter que chaque mot représentatif d'une nuance peut être stocké associé à un coefficient, le coefficient de nuance pour la proposition étant par exemple le produit des coefficients des mots trouvés dans la proposition. Alternativement, le coefficient de nuance pour la proposition peut être la somme des coefficients des mots trouvés dans la proposition.

Le tableau ci-dessous donne deux exemples d'ensembles de coefficients de quelques mots porteurs de nuances, aussi bien dans une composition par produit (colonne de gauche) que par somme (colonne de

droite). On comprendra que l'invention n'est limitée à aucun mode de calcul du coefficient de nuance.

TERME	NUANCE	
	Exemple 1	Exemple 2
Bien plus, beaucoup, énormément	2	+20%
Plus, un peu plus, deux fois plus	1,25	+10%
Peu, moins, un peu moins	0,8	-10%
Très peu, pratiquement pas	0,5	-20%

Coefficient d'orientation d'une thématique

Les coefficients de nuance et de certitude peuvent constituer deux dimensions distinctes de la matrice sémantique, ou être traitées ensemble comme un coefficient d'orientation (« l'orienteur »).

Il est peut être calculé comme le produit des coefficients de certitude et de nuance. En effet, ces deux concepts sont indépendants. La proximité à une thématique peut par exemple être renforcée dans le négatif par une formulation telle que « le véhicule ne comprend certainement pas de moteur », qui correspondra par exemple a un coefficient de -1.75

Le coefficient d'orientation est ainsi typiquement un nombre réel :

< 0 pour la certitude négative

> 0 pour la certitude affirmative

0 pour l'incertitude

Et dont la valeur absolue est

> 1 pour un renforcement

< 1 pour une relativisation

=1 pour une orientation neutre

A l'issue de l'étape (d), la matrice sémantique obtenue a préférentiellement une structure du type

Thème 1	**Thème 2**	**Thème 3**	**Thème i**
Couverture 1	Couverture 2	Couverture 3	Couverture i
Pertinence 1	Pertinence 2	Pertinence 3	Pertinence i
Orienteur 1	Orienteur 2	Orienteur 3	Orienteur i

Composition de matrices sémantiques

Comme expliqué plus haut, un texte est formé de plusieurs phrases formées elles-mêmes de plusieurs propositions. Une matrice sémantique est avantageusement générée pour un anneau pour chaque proposition.

Dans une étape (e0), les matrices sémantiques d'un anneau sont combinées en une matrice globale : est calculé par les moyens de traitement de données 11 pour chaque anneau et chaque thématique identifiée pour au moins une proposition du texte un coefficient global de couverture de la thématique et/ou d'un coefficient global de pertinence de la thématique, et d'au moins un coefficient global d'orientation de la thématique en fonction de l'ensemble desdits coefficients associés une proposition.

Les matrices de deux propositions sont complémentaires si elles portent sur des thèmes différents. La matrice de sens de l'ensemble des deux propositions est constituée de la juxtaposition des deux matrices (puisqu'aucune thématique n'est commune).

Les matrices de deux propositions sont cohérentes si elles portent sur des thèmes communs avec des orienteurs similaires.

Les matrices de deux propositions sont opposées si elles portent sur des thèmes communs avec des orienteurs opposés (de signes différents, i.e. la différence porte sur le coefficient de certitude de la thématique).

Dans le cas général deux matrices A et B portent sur certains thèmes communs et sur d'autres différents. La matrice résultante S est alors composée d'une colonne par thème appartenant à l'une ou l'autre proposition.

Par exemple les règles suivantes peuvent s'appliquer à la composition de deux colonnes pour un même thème :

- un coefficient global de couverture d'une thématique est calculé comme la somme des coefficients de couverture de la thématique

associée à une proposition moins le nombre de mots de référence de la thématique présents dans au moins deux propositions (en d'autres termes il ne faut compter qu'une fois chaque mot. La couverture de la somme est ainsi comprise entre la plus grande des couvertures (cas ou tous les mots de référence de la thématique trouvés dans une proposition sont également dans l'autre), et la somme (cas ou aucun mot de référence n'est commun aux deux couvertures thématiques). Il est à noter que le coefficient global de couverture peut être facilement recalculé comme le nombre *Nmax* de mots de la thématique contenus dans l'ensemble des propositions) ;

- un coefficient global de pertinence d'une thématique est calculé comme la somme des coefficients de pertinence de la thématique associée à une proposition (puisque les occurrences multiples sont prises en compte) ;
- un coefficient global d'orientation d'une thématique est calculé comme la moyenne des coefficients d'orientation de la thématique associés à une proposition pondérés par les coefficients de couverture de la thématique associés. Par exemple, le coefficient global d'orientation du texte S formé des propositions A et B est donné par la formule OS = (OA*CA + OB*CB) / CS

Par ailleurs, on définit la « divergence thématique » comme représentant les variations de sens pour une thématique dans un texte.

Avantageusement, l'étape (e0) comprend ainsi pour chaque thématique le calcul d'un coefficient global de divergence de la thématique. Il se calcule par exemple comme étant l'écart type de la distribution des produits des orienteurs par les couvertures des propositions concernées ramenée au produit holiste de l'orienteur par la couverture du texte global.

Un texte à forte divergence est un texte dans lequel le sujet porté par la thématique est abordé avec des interrogations, des comparaisons, des

confrontations. Un texte à faible divergence est un texte présentant constamment le même angle de vue.

Anneaux sémantiques croissants et décroissants

La notion d'anneau sémantique croissant ou décroissant est relative à un morphisme, permettant de calculer un « coefficient sémantique », représentatif d'un degré de sens porté par le sous-groupe en fonction desdits coefficients de couverture, pertinence et/ou orientation de la thématique, en particulier les coefficients globaux.

Ce coefficient est calculé par les moyens de traitement de données à l'étape (e) du procédé.

Par exemple, soit M le morphisme de $\mathbf{P}(P) \rightarrow R$ tel que

$\forall A \in \mathbf{P}(P)$, avec $T \in \mathbf{P}(V)$, M(A,T) = pertinence(A,T) * orienteur(A,T) * $\sqrt{[1 + \text{divergence}(A,T)^2]}$

M(A,T) est le coefficient sémantique de l'anneau A de la proposition P par rapport à la thématique T selon le vocabulaire V.

M(A) est le coefficient sémantique de l'anneau A de la proposition P par rapport à toutes les thématiques selon le vocabulaire V.

Alternativement, sont possibles (en particulier dans un mode de réalisation ne comprenant pas le calcul d'un coefficient de divergence) des morphismes M tels que

$\forall A \in \mathbf{P}(P)$, avec $T \in \mathbf{P}(V)$, $M(A,T) = [\text{pertinence}(A,T)]^2$ * orienteur(A,T), ou encore

$\forall A \in \mathbf{P}(P)$, avec $T \in \mathbf{P}(V)$, M(A,T) = pertinence(A,T) * couverture(A,T)

Toutes ces formules peuvent également être normalisées.

Quelque soit le morphisme choisi, le coefficient sémantique permet de sélectionner des couples anneaux/thématique les plus porteurs de sens dans une étape (f). En particulier, ce peut être ceux pour lesquels le coefficient est le plus élevé, mais alternativement on peut utiliser le critère de « croissance » des anneaux sémantiques.

On appelle anneau sémantique croissant selon M, tout élément A de **P**(Q) pour lequel :

- $\forall A' \in \mathbf{P}(A)$,
- $\exists T, M(A,T) > M(A',T) + M(A'\Delta A,T)$
- Avec cardinalité(A) > 1

En d'autres termes, un anneau sémantique croissant est un anneau porteur d'un sens plus grand que la somme des sens de ses parties. Pour reformuler encore, il existe une thématique telle que pour toute partition de l'anneau, la somme des coefficients sémantiques des parties de la partition de l'anneau par rapport à cette thématique est inférieure au coefficient sémantique de l'anneau entier par rapport à cette thématique.

Par opposition, les autres anneaux sémantiques sont dit décroissants.

Avantageusement, les couples sous-ensemble/thématique sélectionnés à l'étape (f) sont ceux pour lesquels l'anneau est croissant pour cette thématique.

Le choix du morphisme est déterminant pour sélectionner les anneaux sémantiques. Un morphisme trop lâche conduira à ce que tous les anneaux soit des anneaux sémantiques croissants. Un morphisme trop strict conduira à l'absence d'anneaux sémantiques croissants.

Pour illustrer cette notion d'anneaux croissants/décroissants, dans la proposition « ce véhicule est grand dedans et petit dehors », les anneaux {véhicule, grand} et {véhicule, petit} portent plus de sens que l'anneau global {véhicule, grand, petit}, puisque la présence simultanée des termes grand et petit fait baisser l'orienteur. L'anneau {véhicule, grand, petit} est donc décroissant.

Dans la proposition : « ce véhicule est grand et bleu », les anneaux {véhicule, grand} et {véhicule, bleu} portent moins de sens que l'anneau global {véhicule, grand, bleu}. Ce dernier est croissant.

L'union de deux anneaux sémantiques décroissants est un anneau sémantique décroissant. L'union d'un anneau sémantique décroissant et d'un anneau sémantique croissant est un anneau sémantique décroissant.

L'union de deux anneaux sémantiques croissants est un anneau sémantique soit croissant, soit décroissant. Le caractère croissant est récessif vis-à-vis de l'union.

Un anneau sémantique expressif est un ensemble de mots porteur d'un sens culturel supérieur à celui de l'union de ses parties.

Par exemple dans l'expression : « ce véhicule est une vraie bombe », l'anneau expressif {véhicule, bombe} associée à une nuance de renforcement (« vraie ») porte un sens expressif non présent dans les anneaux singletons {véhicule} et {bombe} et non présent dans l'anneau décroissant {véhicule, bombe}.

Un anneau expressif A est un anneau décroissant devenu croissant par un renforcement de nuance (i.e. grâce à un coefficient de nuance élevé dû à la présence de « vraie » entrainant un orienteur élevé). Le morphisme M présente alors une discontinuité au voisinage de A.

Il est à noter qu'avant même la mise en œuvre de l'étape (f), certains filtres peuvent éliminer certains anneaux selon un paramétrage du moteur.

Il est à noter qu'une notion de connexité entre anneaux et thématiques peut être surveillée par les moyens de traitement de données 11. Un anneau fortement connexe à une thématique sera toujours sélectionné en couple avec cette thématique et jamais une autre (voir plus loin).

Classification du texte

Un schéma global du procédé d'analyse sémantique selon l'invention est représenté par la **figure 2**.

La première partie, qui correspond aux étapes (a) à (f) déjà décrite, est mise en œuvre par un bloc appelé l'analyseur permettant de sélectionner les couples anneaux/thématiques représentatifs du sens du texte.

Dans une étape (g), un classificateur associe les catégories aux textes à l'aide des anneaux sélectionnés. En particulier, les catégories

correspondent à des groupes de couples sous-ensemble/thématique de référence sont stockés sur les moyens de stockage de données 12, et les catégories dans lesquelles le texte est classifié sont celles comprenant au moins un couple sous-ensemble/thématique sélectionné à l'étape (f).

D'autres paramètres peuvent contribuer à la classification, telle que la « dilution ». L'étape (g) peut ainsi comprendre le calcul d'un coefficient dit de dilution, qui représente le nombre d'occurrences de termes des thématiques liées à la ou les catégories déterminées (en d'autres termes les thématiques des couples des groupes associés aux catégories), présents dans le texte rapporté au nombre total de termes desdites thématiques. On dit alors que le texte est de catégorie X selon la dilution D.

Dans un souci d'optimisation, une estimation de ces paramètres et notamment du coefficient de dilution peut être plus précoce dans le procédé.

Apprentissage et enrichissement

Comme expliqué, les catégories ne sont pas figées et peuvent évoluer. En particulier de nouvelles catégories peuvent être générées et d'autres segmentées.

Si aucune catégorie n'est retenue, une nouvelle catégorie pourra être générée portant un sens nouveau : un nouveau groupe est créé si aucun groupe de couples sous-ensemble/thématique de référence ne contient au moins un couple sous-ensemble/thématique sélectionné pour le texte. Les couples sous-ensemble/thématique deviennent ceux de référence de ce groupe.

Lorsqu'une catégorie devient trop peuplée, une segmentation paramétrable la scinde en deux ou plusieurs catégories.

Par ailleurs, les anneaux de propositions non traités par la classification et répondant à certains critères (de score) peuvent être placés dans une pile d'attente.

Ainsi, chaque couple sous-ensemble/thématique de référence peut être associé à un score stocké sur les moyens de stockage de données 12, le score d'un couple sous-ensemble/thématique de référence diminuant avec le temps (par exemple suivant un amortissement hyperbolique) mais augmentant à chaque fois que ce couple sous-ensemble/thématique est sélectionné pour un texte.

En d'autres termes, l'enrichissement repose sur deux mécanismes simultanés :

- Le « score » d'un couple anneau/thématique augmente à chaque fois qu'un même anneau est issu de l'analyse
- Le score d'un couple anneau/thématique s'érode avec le temps selon un amortissement hyperbolique.

Et le procédé peut alors comprendre une étape (h) de suppression d'un couple sous-ensemble/thématique de référence d'un groupe si le score dudit couple passe en dessous d'un premier seuil, ou de modification sur les moyens de stockage de données 12 de ladite pluralité de listes associées aux thématiques si le score dudit couple passe au-dessus d'un deuxième seuil.

En particulier, si le score dépasse le deuxième seuil, plusieurs cas peuvent se présenter selon la « connexité » entre l'anneau et la thématique, comme évoqué précédemment.

La connexité entre un anneau et une thématique peut en effet être représentée par un coefficient représentant pour chaque thématique la fréquence d'apparition de cette thématique parmi les thématiques telles que le couple anneau/thématique associé a déjà été sélectionné. En d'autres termes la connexité entre un anneau et une thématique est par exemple donnée comme le score de ce couple anneau/thématique sur la somme des scores associés à des couples de cet anneau avec une thématique de référence.

Les différents cas qui peuvent se présenter sont :

- les anneaux non connexes aux thématiques donnent naissance à de nouvelles thématiques (création d'une nouvelle thématique

pour laquelle la liste de mot associée est définie par l'anneau du couple dont le score a dépassé le deuxième seuil) ;

- les anneaux fortement connexes à une thématique (par exemple connexité supérieure à 90%) sont fusionnés dans la thématique connexe (par exemple, si un anneau est très proche d'une thématique mais comprend un mot de plus, ce mot finit par être ajouté à la liste de mots associée à la thématique).

A l'inverse, un anneau fortement érodé (score passant en-dessous du premier seuil) disparait de la pile. Les deux seuils peuvent être définis manuellement en fonction de la « sensibilité », c'est-à-dire le niveau souhaité d'évolutivité du système. Des seuils proches (premier seuil élevé et/ou deuxième seuil bas) entrainent un fort renouvellement des thématiques et catégories.

REVENDICATIONS

1. Procédé d'analyse sémantique d'un texte en langage naturel reçu par un équipement (1) depuis des moyens de saisie (14), le procédé étant caractérisé en ce qu'il comprend la mise en œuvre par des moyens de traitement de données (11) de l'équipement (1) d'étapes de :

(a) Découpage syntaxique d'au moins une partie du texte en une pluralité de mots ;

(b) Filtrage des mots de ladite partie de texte par rapport à une pluralité de liste de mots de référence stockées sur des moyens de stockage de données (12) de l'équipement (1), chacune étant associée à une thématique, de sorte à identifier :

- L'ensemble des mots de ladite partie du texte associés à au moins une thématique,
- L'ensemble des thématiques de ladite partie du texte ;

(c) Construction d'une pluralité de sous-ensembles de l'ensemble des mots de ladite partie du texte associés à au moins une thématique ;

(d) Pour chacun desdits sous-ensembles et pour chaque thématique identifiée, calcul :

- d'un coefficient de couverture de la thématique et/ou d'un coefficient de pertinence de la thématique en fonction d'occurrences dans ladite partie du texte de mots de référence associés à la thématique ;
- d'au moins un coefficient d'orientation de la thématique à partir des mots de ladite partie du texte ne faisant pas partie du sous-ensemble ;

(e) Pour chacun desdits sous-ensembles et pour chaque thématique identifiée, calcul d'un coefficient sémantique représentatif d'un degré de sens porté par le sous-groupe en fonction desdits coefficients de couverture, pertinence et/ou orientation de la thématique.

(f) Sélection en fonction des coefficients sémantiques d'au moins un couple sous-ensemble/thématique.

(g) Classification du texte en fonction dudit au moins un couple sous-ensemble/thématique sélectionné.

2. Procédé selon la revendication 1, dans lequel un coefficient de couverture d'une thématique est calculé à l'étape (d) comme le nombre *N* de mots de référence associés à la thématique compris dans ledit sous-ensemble.

3. Procédé selon l'une des revendications précédentes, dans lequel un coefficient de pertinence d'une thématique est calculé à l'étape (d) par la formule $N * (1 + \ln(R))$, où *N* est le nombre de mots de référence associés à la thématique compris dans le sous-ensemble et *R* le nombre total d'occurrences dans ladite partie du texte de mots de référence associés à la thématique.

4. Procédé selon l'une des revendications précédentes, dans lequel deux coefficients d'orientation de la thématique sont calculés à l'étape (c), dont un coefficient de certitude de la thématique et un coefficient de nuance de la thématique.

5. Procédé selon la revendication 4, dans lequel un coefficient de certitude d'une thématique est calculé à l'étape (d) comme valant :

- 1 si les mots ne faisant pas partie du sous-ensemble sont représentatifs d'une proximité affirmative avec la thématique ;
- -1 si les mots ne faisant pas partie du sous-ensemble sont représentatifs d'une proximité négative avec la thématique ;
- 0 si les mots ne faisant pas partie du sous-ensemble sont représentatifs d'une proximité incertaine avec la thématique.

6. Procédé selon l'une des revendications 4 et 5, dans lequel un coefficient de nuance d'une thématique est un scalaire positif supérieur à 1 lorsque les mots ne faisant pas partie du sous-ensemble sont représentatifs d'une amplification de la thématique, et un scalaire positif inférieur à 1 lorsque les mots ne faisant pas partie du sous-ensemble sont représentatifs d'une atténuation de la thématique.

7. Procédé selon l'une des revendications précédentes, comprenant une étape (a0) préalable de découpage du texte en une pluralité de proposition, chacune étant une partie du texte pour laquelle les étapes (a) à (d) du procédé selon répétées de sorte à obtenir pour chaque proposition un ensemble de coefficients de couverture, de pertinence, et/ou d'orientation associés à la proposition, le procédé comprenant préalablement à l'étape (e) une étape (e0) de calcul pour chacun desdits sous-ensemble et pour chaque thématique identifiée pour au moins une proposition du texte d'un coefficient global de couverture de la thématique et/ou d'un coefficient global de pertinence de la thématique, et d'au moins un coefficient global d'orientation de la thématique en fonction de l'ensemble desdits coefficients associés une proposition.

8. Procédé selon la revendication 7, dans lequel un coefficient global de couverture d'une thématique est calculé à l'étape (e0) comme la somme des coefficients de couverture de la thématique associée à une proposition moins le nombre de mots de référence de la thématique présents dans au moins deux propositions.

9. Procédé selon l'une des revendications 7 et 8, dans lequel un coefficient global de pertinence d'une thématique est calculé à l'étape (e0) comme la somme des coefficients de pertinence de la thématique associée à une proposition.

10. Procédé selon l'une des revendications 7 à 9, dans lequel un coefficient global d'orientation d'une thématique est calculé à l'étape (e0) comme la moyenne des coefficients d'orientation de la thématique associés à une proposition pondérés par les coefficients de couverture de la thématique associés.

11. Procédé selon l'une des revendications 7 à 10, dans lequel l'étape (e0) comprend pour chacun desdits sous-ensembles et pour chaque thématique le calcul d'un coefficient global de divergence de la thématique correspondant à l'écart-type de la distribution des produits des coefficients d'orientation par les coefficients de couverture associés à chaque proposition.

12. Procédé selon la revendication 11, dans lequel un coefficient sémantique d'un sous-ensemble *A* pour une thématique *T* est calculé à l'étape (e) par la formule M(A,T) = coefficient de pertinence(A,T) * coefficient d'orientation(A,T) * $\sqrt{}$ [1 + coefficient de divergence(A,T)²].

13. Procédé selon l'une des revendications précédentes, dans lequel les couples sous-ensemble/thématique sélectionnés à l'étape (f) sont ceux tels que pour toute partition du sous-ensemble en une pluralité de parties dudit sous-ensemble, le coefficient sémantique du sous-ensemble pour la thématique est supérieur à la somme des coefficients sémantiques des sous-parties du sous-ensemble pour la thématique.

14. Procédé selon l'une des revendications précédentes, dans lequel des groupes de couples sous-ensemble/thématique de référence sont stockés sur les moyens de stockage de données (12), l'étape (g) comprenant la détermination du ou des groupes comprenant au moins un couple sous-ensemble/thématique sélectionné à l'étape (f).

15. Procédé selon la revendication 14, dans lequel l'étape (g) comprend la création d'un nouveau groupe si aucun groupe de couples

sous-ensemble/thématique de référence ne contient au moins un couple sous-ensemble/thématique sélectionné pour le texte.

16. Procédé selon l'une des revendications 14 et 15, dans lequel chaque couple sous-ensemble/thématique de référence est associé à un score stocké sur les moyens de stockage de données (12), le score d'un couple sous-ensemble/thématique de référence diminuant avec le temps mais augmentant à chaque fois que ce couple sous-ensemble/thématique est sélectionné pour un texte.

17. Procédé selon la revendications 16, comprenant une étape (h) de suppression d'un couple sous-ensemble/thématique de référence d'un groupe si le score dudit couple passe en dessous d'un premier seuil, ou de modification sur les moyens de stockage de données (12) de ladite pluralité de listes associées aux thématiques si le score dudit couple passe au-dessus d'un deuxième seuil.

18. Procédé selon l'une des revendications 14 à 17, dans lequel l'étape (g) comprend pour chaque groupe de couples sous-ensemble/thématique de référence le calcul d'un coefficient de dilution représentant le nombre d'occurrences dans ladite partie du texte de mots de référence associés à des thématiques des couples sous-ensemble/thématique de référence présents dans le texte rapporté au nombre total de mots de référence associés auxdites thématiques.

19. Procédé selon l'une des revendications précédentes, dans lequel tous les sous-ensembles de l'ensemble des mots de ladite partie du texte associés à au moins une thématique sont construits à l'étape (c).

20. Equipement (1) comprenant des moyens de traitement de données (11) configurés pour mettre en œuvre suite à la réception d'un

texte en langage naturel un procédé selon l'une des revendication précédentes d'analyse sémantique du texte.

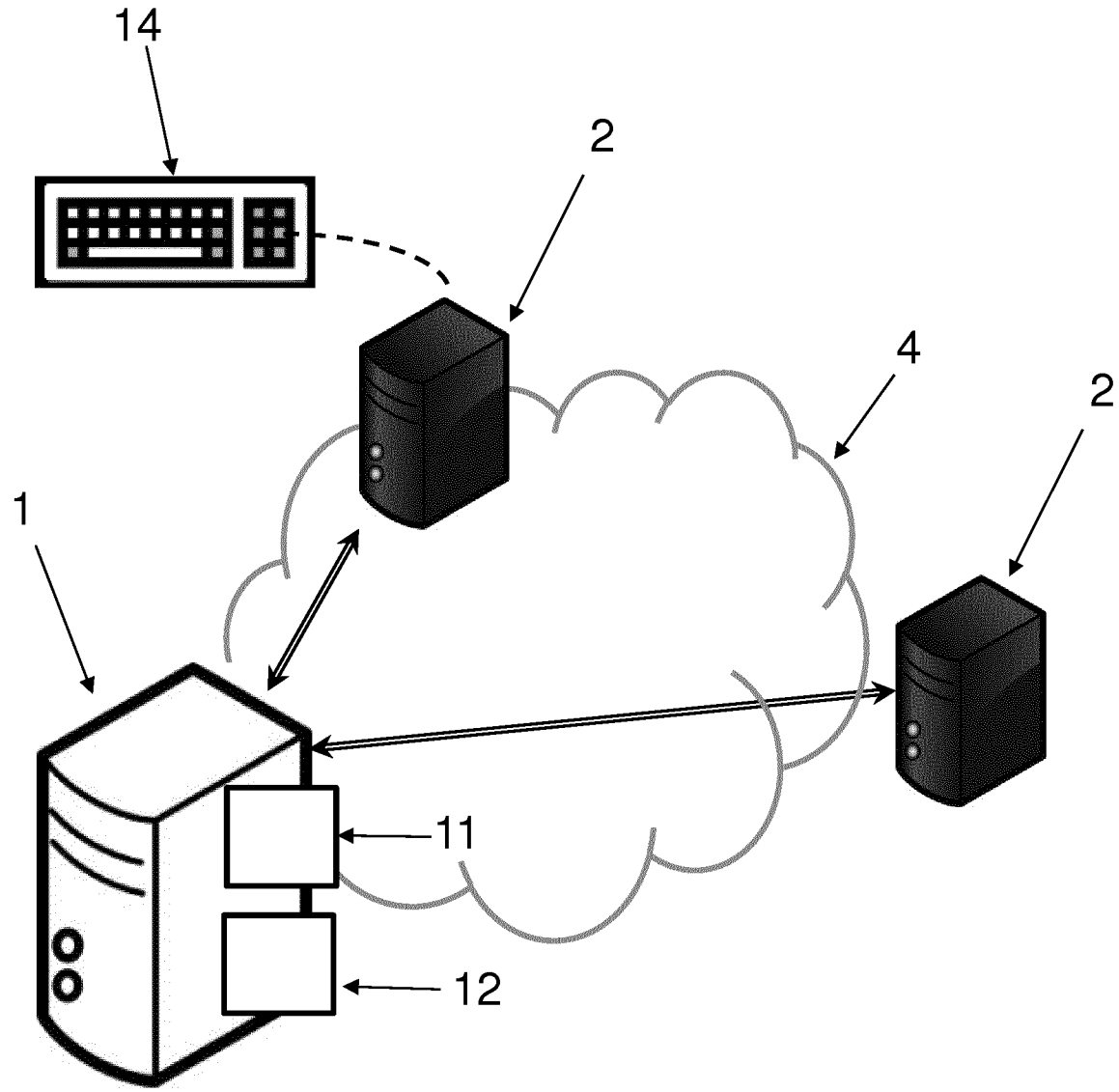

FIG. 1

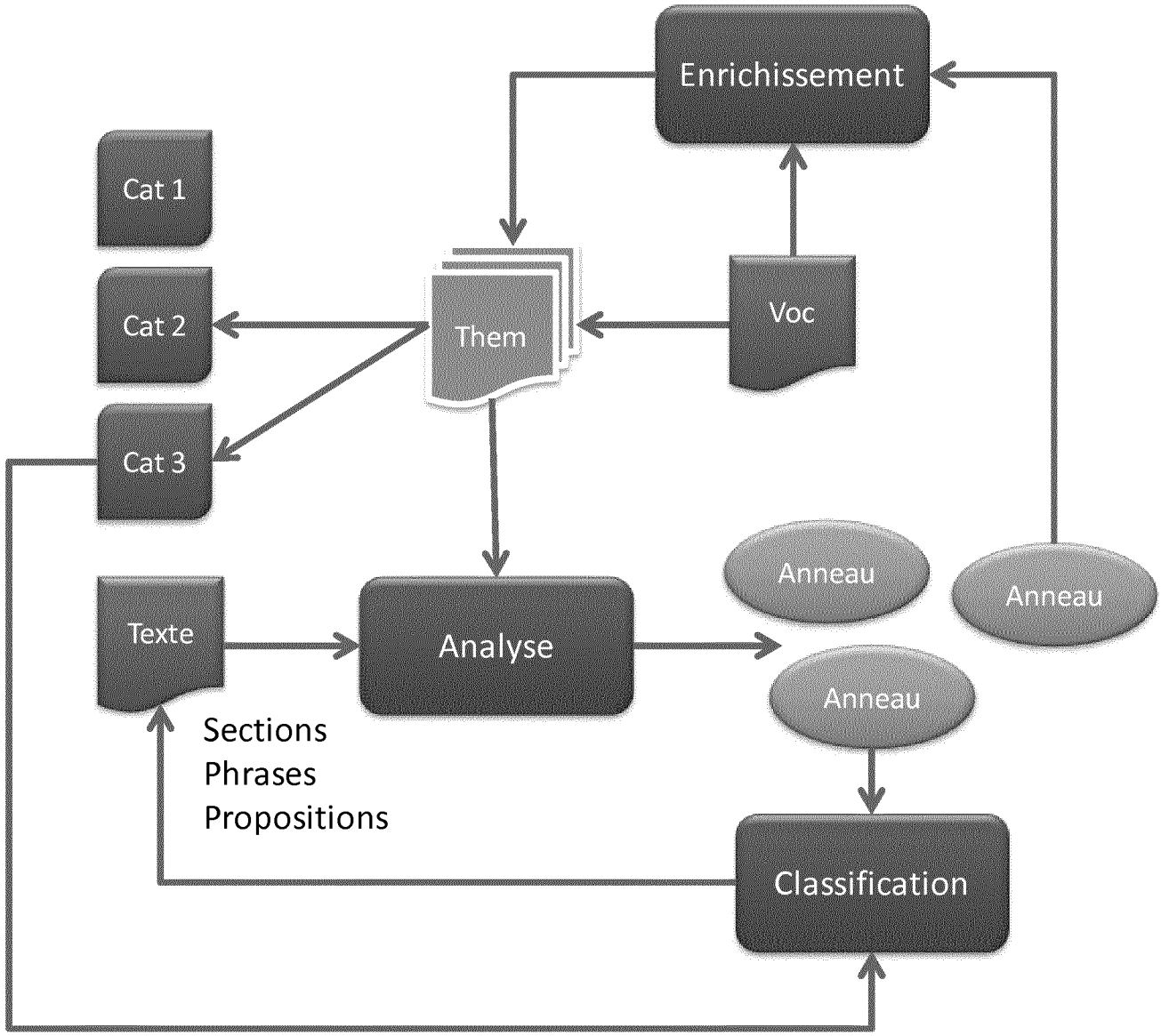
Enrichissement
Cat 1
Cat 2
Cat 3
Them
Voc
Texte
Analyse
Anneau
Anneau
Anneau
Sections
Phrases
Propositions
Classification

FIG. 2

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2015/051722

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06F17/27 G06F17/30
ADD.

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 practicable, search terms used)
EPO-Internal, WPI Data, COMPENDEX, INSPEC, IBM-TDB

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	EP 1 650 680 A2 (HEWLETT PACKARD DEVELOPMENT CO [US]) 26 April 2006 (2006-04-26) abstract paragraph [0017] - paragraph [0037]; figures 1,2 paragraph [0050] - paragraph [0068]; figure 3	1-20

A	GRIGORI SIDOROV ET AL: "Syntactic Dependency-Based N-grams as Classification Features", 27 October 2012 (2012-10-27), ADVANCES IN COMPUTATIONAL INTELLIGENCE, SPRINGER BERLIN HEIDELBERG, BERLIN, HEIDELBERG, PAGE(S) 1 - 11, XP047026724, ISBN: 978-3-642-37797-6 the whole document	1-20

	-/--	

[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 application or patent 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
25 February 2015	10/03/2015

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	Lechenne, Laurence

1

Form PCT/ISA/210 (second sheet) (April 2005)

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	APOORV AGARWAL ET AL: "Contextual phrase-level polarity analysis using lexical affect scoring and syntactic N-grams", PROCEEDINGS OF THE 12TH CONFERENCE OF THE EUROPEAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS ON, EACL '09, 30 March 2009 (2009-03-30), pages 24-32, XP055162964, Morristown, NJ, USA DOI: 10.3115/1609067.1609069 abstract section 1.Introduction; page 24 - page 25	1-20
A	US 2011/179032 A1 (CEUSTERS WERNER [BE] ET AL) 21 July 2011 (2011-07-21) abstract paragraph [0109] - paragraph [0123]; figures 5-7 paragraph [0157] - paragraph [0183]; figure 10	1-20
A	EP 1 462 950 A1 (SONY INT EUROPE GMBH [DE] SONY DEUTSCHLAND GMBH [DE]) 29 September 2004 (2004-09-29) the whole document	1-20
A	T Revathi ET AL: "Sentence Level Semantic Classification of Online Product Reviews of Mixed Opinions Using Naive bayes Classifier", International Journal of Engineering Trends and Technology, 1 January 2012 (2012-01-01), pages 3-2, XP055162614, Retrieved from the Internet: URL:http://www.ijettjournal.org/volume-3/issue-2/IJETT-V3I2P213.pdf [retrieved on 2015-01-16] the whole document	1-20

1

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/EP2015/051722

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
EP 1650680 A2	26-04-2006	EP 1650680 A2	26-04-2006
		JP 4713870 B2	29-06-2011
		JP 2006113746 A	27-04-2006
		KR 20060052194 A	19-05-2006
		US 2006112040 A1	25-05-2006
US 2011179032 A1	21-07-2011	US 7493253 B1	17-02-2009
		US 2009259459 A1	15-10-2009
		US 2011179032 A1	21-07-2011
		US 2013185303 A1	18-07-2013
		US 2013211823 A1	15-08-2013
EP 1462950 A1	29-09-2004	DE 60315947 T2	21-05-2008
		EP 1462950 A1	29-09-2004

RAPPORT DE RECHERCHE INTERNATIONALE

Demande internationale n°
PCT/EP2015/051722

A. CLASSEMENT DE L'OBJET DE LA DEMANDE
INV. G06F17/27 G06F17/30
ADD.

Selon la classification internationale des brevets (CIB) ou à la fois selon la classification nationale et la CIB

B. DOMAINES SUR LESQUELS LA RECHERCHE A PORTE

Documentation minimale consultée (système de classification suivi des symboles de classement)
G06F

Documentation consultée autre que la documentation minimale dans la mesure où ces documents relèvent des domaines sur lesquels a porté la recherche

Base de données électronique consultée au cours de la recherche internationale (nom de la base de données, et si cela est réalisable, termes de recherche utilisés)
EPO-Internal, WPI Data, COMPENDEX, INSPEC, IBM-TDB

C. DOCUMENTS CONSIDERES COMME PERTINENTS

Catégorie*	Identification des documents cités, avec, le cas échéant, l'indication des passages pertinents	no. des revendications visées
X	EP 1 650 680 A2 (HEWLETT PACKARD DEVELOPMENT CO [US]) 26 avril 2006 (2006-04-26) abrégé alinéa [0017] - alinéa [0037]; figures 1,2 alinéa [0050] - alinéa [0068]; figure 3 -----	1-20
A	GRIGORI SIDOROV ET AL: "Syntactic Dependency-Based N-grams as Classification Features", 27 octobre 2012 (2012-10-27), ADVANCES IN COMPUTATIONAL INTELLIGENCE, SPRINGER BERLIN HEIDELBERG, BERLIN, HEIDELBERG, PAGE(S) 1 - 11, XP047026724, ISBN: 978-3-642-37797-6 le document en entier -----	1-20

-/--

[X] Voir la suite du cadre C pour la fin de la liste des documents

[X] Les documents de familles de brevets sont indiqués en annexe

* Catégories spéciales de documents cités:

"A" document définissant l'état général de la technique, non considéré comme particulièrement pertinent

"E" document antérieur, mais publié à la date de dépôt international ou après cette date

"L" document pouvant jeter un doute sur une revendication de priorité ou cité pour déterminer la date de publication d'une autre citation ou pour une raison spéciale (telle qu'indiquée)

"O" document se référant à une divulgation orale, à un usage, à une exposition ou tous autres moyens

"P" document publié avant la date de dépôt international, mais postérieurement à la date de priorité revendiquée

"T" document ultérieur publié après la date de dépôt international ou la date de priorité et n'appartenant pas à l'état de la technique pertinent, mais cité pour comprendre le principe ou la théorie constituant la base de l'invention

"X" document particulièrement pertinent; l'invention revendiquée ne peut être considérée comme nouvelle ou comme impliquant une activité inventive par rapport au document considéré isolément

"Y" document particulièrement pertinent; l'invention revendiquée ne peut être considérée comme impliquant une activité inventive lorsque le document est associé à un ou plusieurs autres documents de même nature, cette combinaison étant évidente pour une personne du métier

"&" document qui fait partie de la même famille de brevets

Date à laquelle la recherche internationale a été effectivement achevée	Date d'expédition du présent rapport de recherche internationale
25 février 2015	10/03/2015

Nom et adresse postale de l'administration chargée de la recherche internationale	Fonctionnaire autorisé
Office Européen des Brevets, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Lechenne, Laurence

1

C(suite). DOCUMENTS CONSIDERES COMME PERTINENTS

Catégorie*	Identification des documents cités, avec, le cas échéant, l'indication des passages pertinents	no. des revendications visées
A	APOORV AGARWAL ET AL: "Contextual phrase-level polarity analysis using lexical affect scoring and syntactic N-grams", PROCEEDINGS OF THE 12TH CONFERENCE OF THE EUROPEAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS ON, EACL '09, 30 mars 2009 (2009-03-30), pages 24-32, XP055162964, Morristown, NJ, USA DOI: 10.3115/1609067.1609069 abrégé section 1.Introduction; page 24 - page 25	1-20
A	US 2011/179032 A1 (CEUSTERS WERNER [BE] ET AL) 21 juillet 2011 (2011-07-21) abrégé alinéa [0109] - alinéa [0123]; figures 5-7 alinéa [0157] - alinéa [0183]; figure 10	1-20
A	EP 1 462 950 A1 (SONY INT EUROPE GMBH [DE] SONY DEUTSCHLAND GMBH [DE]) 29 septembre 2004 (2004-09-29) le document en entier	1-20
A	T Revathi ET AL: "Sentence Level Semantic Classification of Online Product Reviews of Mixed Opinions Using Naive bayes Classifier", International Journal of Engineering Trends and Technology, 1 janvier 2012 (2012-01-01), pages 3-2, XP055162614, Extrait de l'Internet: URL:http://www.ijettjournal.org/volume-3/issue-2/IJETT-V3I2P213.pdf [extrait le 2015-01-16] le document en entier	1-20

RAPPORT DE RECHERCHE INTERNATIONALE

Renseignements relatifs aux membres de familles de brevets

Demande internationale n°
PCT/EP2015/051722

Document brevet cité au rapport de recherche	Date de publication	Membre(s) de la famille de brevet(s)	Date de publication
EP 1650680 A2	26-04-2006	EP 1650680 A2	26-04-2006
		JP 4713870 B2	29-06-2011
		JP 2006113746 A	27-04-2006
		KR 20060052194 A	19-05-2006
		US 2006112040 A1	25-05-2006
US 2011179032 A1	21-07-2011	US 7493253 B1	17-02-2009
		US 2009259459 A1	15-10-2009
		US 2011179032 A1	21-07-2011
		US 2013185303 A1	18-07-2013
		US 2013211823 A1	15-08-2013
EP 1462950 A1	29-09-2004	DE 60315947 T2	21-05-2008
		EP 1462950 A1	29-09-2004