(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
15 août 2013 (15.08.2013)

WIPO | PCT

(10) Numéro de publication internationale
WO 2013/117872 A1

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

(21) Numéro de la demande internationale : PCT/FR2013/050269

(22) Date de dépôt international : 8 février 2013 (08.02.2013)

(25) Langue de dépôt : français

(26) Langue de publication : français

(30) Données relatives à la priorité :
1251241 9 février 2012 (09.02.2012) FR

(71) Déposant : **MINING ESSENTIAL** [FR/FR]; 85 Rue Camille Groult, F-94400 Vitry Sur Seine (FR).

(72) Inventeur : **LEHMAM, Abderrafih**; Mining Essential, 85 Rue Camille Groult, F-94400 Vitry Sur Seine (FR).

(74) Mandataires : **CAMUS, Olivier** et al.; Cabinet CAMUS LEBKIRI, 87 Rue Taitbout, F-75009 Paris (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, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, 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, ML, MR, NE, SN, TD, TG).

[Suite sur la page suivante]

(54) Title : METHOD FOR IDENTIFYING A SET OF SENTENCES IN A DIGITAL DOCUMENT, METHOD FOR GENERATING A DIGITAL DOCUMENT, AND ASSOCIATED DEVICE

(54) Titre : PROCEDE D'IDENTIFICATION D'UN ENSEMBLE DE PHRASES D'UN DOCUMENT NUMERIQUE, PROCEDE DE GENERATION D'UN DOCUMENT NUMERIQUE, DISPOSITIF ASSOCIE

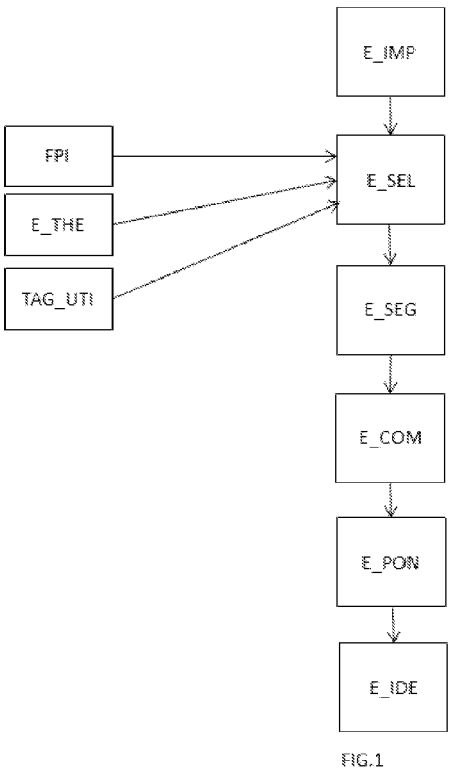


FIG.1

(57) Abstract : The invention relates to a method for generating a digital document, known as a "digital summary", said method comprising: a parameterisation step for defining a first degree of summarisation of a first digital document defining a first ratio between a first number representing the quantity of data contained in the desired digital abstract and a second number representing the quantity of data contained in the first document; an analysis step for analysing the first digital document, comprising the definition of a set of terms, known as TAG; a segmentation step for (i) determining a first set of sentences in the first document or (ii) associating a weighting with each of the sentences; an extraction step for extracting a number of sentences according to the degree of condensation; and a generation step for generating a digital abstract comprising a set of ordered sentences.

(57) Abrégé : Le procédé de génération d'un document numérique, noté « résumé numérique », comprend : une étape de paramétrage permettant de définir un premier degré de condensation d'un premier document numérique définissant

[Suite sur la page suivante]

WO 2013/117872 A1

Publiée :

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

un rapport entre un premier nombre représentant la quantité de données contenues dans le résumé numérique souhaité et un second nombre représentant la quantité de données contenues dans le premier document; · une étape d'analyse du premier document numérique comprenant la définition d'un ensemble de termes, noté TAG; · une étape de segmentation permettant de : o déterminer un premier ensemble de phrases du premier document; o associer une pondération à chacune des phrases · une étape d'extraction d'un nombre de phrases selon le degré de condensation; · une étape de génération d'un résumé numérique comprenant un ensemble de phrases ordonnées.

PROCEDE D'IDENTIFICATION D'UN ENSEMBLE DE PHRASES D'UN DOCUMENT NUMERIQUE, PROCEDE DE GENERATION D'UN DOCUMENT NUMERIQUE, DISPOSITIF ASSOCIE

DOMAINE

L'invention concerne le domaine des procédés et des systèmes permettant d'extraire des données pertinentes et exploitables selon certains critères d'un corpus de documents numériques. Plus particulièrement, le domaine de l'invention se rapporte aux procédés de génération d'un résumé d'un document numérique dont certaines caractéristiques sont paramétrables.

ETAT DE L'ART

Actuellement certains procédés permettent, à partir d'un document numérique, d'identifier des passages ou des extraits de ce document à partir d'une méthode statistique. Ces méthodes visent à extraire des données d'un document numérique, par exemple des mots ou des phrases, en fonction d'occurrences de certains TAGS prédéfinis dans le document.

Les méthodes actuelles qui permettent de générer dynamiquement un résumé d'un document numérique ne semblent pas proposer un niveau de cohérence et de fidélité suffisant pour être exploitables par un utilisateur.

En effet, une difficulté de telles méthodes est de permettre à un utilisateur d'accéder aux éléments essentiels d'un document numérique par le bais de la génération d'un résumé. Ce dernier doit comporter une cohérence et une fidélité suffisante pour être exploitable. Les méthodes actuelles se basent sur une sémantique définie par un utilisateur, par la définition de mots clefs par exemple, qui ne suffit pas à elle seule à conserver une cohérence et un sens du document numérique. Il est même possible par l'utilisation de telles méthodes de dénaturer la cohérence d'un document numérique ou de générer un contre sens en décontextualisant certaines données du document numérique.

RESUME DE L'INVENTION

L'invention permet de résoudre les inconvénients précités.

L'invention a pour objet un procédé d'identification d'un ensemble de phrases d'un premier document numérique. Le procédé d'identification comprend :

- une étape d'importation du premier document numérique dans au moins un format prédéfini permettant : soit d'afficher le document dans une première interface soit de le stocker dans une mémoire ;
- une étape de sélection dans une base de fragments de phrases indicateurs, notée FPI dont chacun des termes peut être décliné grâce à un dictionnaire morphologique, ladite FPI comprenant un ensemble de TAG linguistiques, chacun des TAG linguistiques comprenant une première attribution de valeurs numériques choisies dans un premier intervalle défini par une première valeur minimale et une première valeur maximale ;
- une étape de segmentation du premier document numérique permettant de :
  - déterminer un premier ensemble de phrases du premier document ;
  - numéroter les phrases de ce premier ensemble définissant une première séquence ;
- une étape de comparaison des termes de chaque phrase du premier document segmenté et des TAG linguistiques de la base de fragments de phrases indicateurs permettant de repérer la présence des TAG linguistiques dans lesdites phrases ;
- une étape de pondération de chacune des phrases par attribution d'un premier score correspondant à la somme des valeurs de chaque TAG linguistique repérés dans chacune des phrases ;
- une étape d'identification d'un second ensemble de phrases compris dans le premier ensemble de phrases ayant une pondération supérieure à un premier seuil.

Dans un mode amélioré, le procédé d'identification d'un ensemble de phrases d'un premier document numérique :

- l'étape de sélection comprend la sélection d'un thésaurus définissant un fichier comprenant une liste de TAG sémantiques d'un domaine, chacun des TAG sémantiques comprenant une seconde attribution de valeurs pour chaque TAG sémantique comprises dans un second intervalle définit par une seconde valeur minimale et une seconde valeur maximale;
- l'étape de pondération de chacune des phrases par attribution d'un second score correspondant à la somme des valeurs de chaque TAG sémantiques repérés dans chacune des phrases.

Dans un autre mode de réalisation qui peut se combiner avec le précédent,

- l'étape de sélection comprend la sélection d'un ensemble de TAG définis par un utilisateur définissant des TAG utilisateurs comprend des expressions sémantiques et/ou des termes, chacun des TAG utilisateurs comprenant une troisième attribution de valeurs pour chaque TAG utilisateurs comprises dans un troisième intervalle définit une troisième valeur minimale et une troisième valeur maximale;
- l'étape de pondération de chacune des phrases par attribution d'un troisième score correspondant à la somme des valeurs de chaque TAG utilisateurs repérés dans chacune des phrases.

Un avantage technique des caractéristiques de l'invention est que la base de fragments de phrases indicateurs permet d'identifier des termes ou des expressions qui peuvent comporter des TAG associés à la structure d'un texte et à l'importance de données spécifiques dans un contexte particulier. De tels TAG peuvent être par exemple : « en conclusion », « pour finir », « le plus important », etc.

Un avantage du procédé de l'invention est que les TAGS de la base de fragments de phrases indicateurs sont dissociés des mots clefs définis par un utilisateur susceptibles de l'intéresser. En outre, un thésaurus

peut être associé de manière à identifier des phrases selon un domaine précis, par exemple le domaine économique.

Avantageusement, le premier seuil est calculé à partir d'un taux de condensation défini par le nombre de phrases souhaité par un utilisateur du second ensemble sur le nombre total de phrases du premier ensemble de phrases.

Avantageusement, le premier seuil est calculé à partir d'un taux de condensation défini par le nombre de termes souhaités par un utilisateur du second ensemble de phrases sur le nombre total de termes du premier ensemble de phrases.

Avantageusement, une interface permette de configurer le taux de condensation.

Avantageusement, une étape d'affichage au moyen d'une interface du premier document numérique comprend la génération des phrases identifiées selon une taille de caractère plus importante que les phrases non identifiées.

Avantageusement, l'étape de comparaison (E_COM) comprend la détermination de termes racines des TAG linguistiques du FPI à partir d'un dictionnaire morphologique et la comparaison des déclinaisons des termes racines des TAG linguistiques avec chaque phrase du document numérique.

Avantageusement, l'étape de pondération comprend la somme des premiers, seconds et/ou troisièmes score(s) pour chacune des phrases du document numérique, définissant ainsi un poids sémantique, le poids sémantique de chaque phrase étant comparé à un seuil prédéfini dans l'étape d'identification.

Avantageusement, la valeur moyenne des valeurs de la seconde attribution (ATT2) est dans un intervalle représentant 20% du premier l'intervalle centré sur la valeur moyenne des valeurs de la première attribution.

Cette configuration permet d'obtenir une très bonne pertinence du résumé généré en termes de conservation de la fidélité du sens général du texte d'origine. Les relations définissant le premier et le second intervalles ont une importance quant au résumé qui est généré et de la fidélité du sens du texte d'origine qui est conservé. La configuration décrite ci-dessus résulte

d'une analyse d'un grand nombre d'essais et a permis un réglage optimal cette configuration.

Avantageusement, la valeur moyenne des valeurs de la troisième attribution (ATT3) est dans un intervalle représentant 20% du premier l'intervalle centré sur la valeur moyenne des valeurs de la première attribution.

Cette configuration permet d'obtenir une très bonne pertinence du résumé généré en termes de conservation de la fidélité du sens général du texte d'origine. Les relations définissant le premier et le troisième intervalles ont une importance quant au résumé qui est généré et de la fidélité du sens du texte d'origine qui est conservé. La configuration décrite ci-dessus résulte d'une analyse d'un grand nombre d'essais et a permis un réglage optimal cette configuration.

En outre, l'objet de l'invention concerne un procédé de génération d'un document numérique, noté « résumé numérique », comprenant la génération et l'affichage sur un afficheur du second ensemble de phrases, les dites phrases étant identifiées à partir du procédé d'identification de l'invention, selon une séquence ordonnée par une numérotation croissante.

Avantageusement, le résumé numérique généré comprend des symboles activables, un symbole activable étant associé à chacune des phrases du second ensemble, les phrases du résumé numérique et les symboles activables étant affichées sur un afficheur de manière à ce que les symboles activables soient affichées à proximité des phrases, l'activation d'au moins un symbole activable d'une phrase sélectionnée générant un second résumé numérique, le second résumé numérique comportant des phrases ordonnées dont la numérotation est successive, cet ensemble comportant ladite phrase sélectionnée et un premier ensemble de phrases dont la numérotation précède celle de la phrase sélectionnée et un second ensemble de phrases dont la numérotation succède à celle de la phrase sélectionnée.

Avantageusement, l'activation d'un symbole activable est réalisé au moyen d'un clic de souris pour ordinateur ou d'un survol d'un curseur sur des données activables ou d'un toucher tactile dans une zone comprenant le symbole activable.

Avantageusement, le symbole activable est un caractère alphanumérique.

Avantageusement, le symbole activable est un numéro représentant le numéro de la phrase dans le premier document.

En outre, l'objet de l'invention concerne un procédé de génération d'un document numérique, appelé « synthèse numérique ».

Avantageusement, le procédé de génération d'un résumé numérique est appliqué à un ensemble de documents numériques de manière à générer une pluralité de résumés numériques, ledit procédé comportant une étape de génération d'une synthèse numérique à partir de la définition d'un paramètre, dit de taux de répartition, représentant la quantification des données de chaque résumé numérique présente dans la synthèse et d'un second taux de condensation de chaque résumé numérique, la synthèse numérique comprenant un ensemble de phrases ordonnées et sélectionnées en fonction du taux de répartition et du second taux de condensation de chacun des résumé numérique.

En outre, l'objet de l'invention concerne un dispositif de génération d'un document numérique comportant un afficheur permettant d'afficher au moins un document numérique, un calculateur permettant de mettre en œuvre les étapes du procédé de l'invention. Le dispositif comprend également une interface permettant de paramétrer au moins un premier taux de condensation, un système de commandes permettant de lancer la génération d'un premier résumé du numérique.

Avantageusement, le système de commandes permet de lancer la génération d'un second résumé numérique du premier résumé numérique.

Avantageusement, l'interface comprend une première fenêtre permettant d'afficher un ensemble de documents numériques et une seconde fenêtre permettant d'afficher un ensemble de résumés numériques correspondant au résumé de chaque document de la première fenêtre.

Avantageusement, l'interface comprend des premiers moyens de sélection d'un taux de condensation d'un résumé numérique, des seconds moyens de sélection d'un thésaurus parmi, une liste de thésaurus prédéfinie et des moyens permettant de définir des TAG d'un utilisateur.

BREVES DESCRIPTION DES FIGURES

D'autres caractéristiques et avantages de l'invention ressortiront clairement de la description qui est donnée ci-après, à titre purement indicatif et nullement limitatif, de modes de réalisation faisant références à différentes figures sur lesquelles :

- la figure 1 représente un schéma des principales étapes du procédé de l'invention.

DESCRIPTION

La figure 1 représente les principales étapes du procédé dont notamment :

- une étape d'importation d'un document numérique, notée E_IMP ;
- une étape de sélection d'un ensemble de fichiers ou de données d'une base de données, tels que la base de fragments de phrases indicateurs, noté FPI, un thésaurus noté THE et définissant un champ lexical d'un domaine ou encore une liste de TAG noté TAG_UTI et définie par un utilisateur ;
- une étape de segmentation E_SEG du document numérique en une pluralité de phrases ;
- une étape de comparaison, notée E_COM, de termes ou d'expressions des phrases du document segmenté avec les TAG de chaque fichiers sélectionnés ;
- une étape de pondération, notée E_PON, permettant d'attribuer un score à chaque phrase ;
- une étape d'identification, notée E_IDE, de phrases ayant un score supérieur à un seuil prédéfini ;
- éventuellement le procédé de l'invention comporte une étape de génération d'un résumé numérique, notée E_GEN, comprenant les phrases identifiées à l'étape E_IDE, les phrases étant affichées selon un séquencement prédéfini.

Dans la suite la description de chaque étape du procédé de l'invention est décrite en détail. Des étapes supplémentaires peuvent être

réalisées dans le procédé dans certains modes de réalisation améliorés de l'invention.

Le procédé de l'invention comprend une étape d'identification d'un premier document numérique dont on souhaite extraire un ensemble de phrases selon un certain nombre de critères. Les phrases extraites permettront dans un mode de réalisation de l'invention de générer un résumé, appelé résumé numérique dans la suite de la description.

Le procédé comprend donc l'identification d'un document numérique, l'identification du document numérique pouvant être réalisée de différentes manières. Ce document peut comporter un titre, une date, une langue ou encore une pluralité de langue, un code de référence pouvant servir d'identifiant. En outre, le document peut comporter des données décrivant sa forme telles que son nombre de page, son nombre de mots, sa disposition ou son format. Le document doit être sous forme numérique, c'est-à-dire comprenant au moins un ensemble de caractères alphanumériques identifiables, par exemple par un logiciel de traitement de texte ou un navigateur internet. Tout type de format du document numérique est compatible du procédé de l'invention à savoir par exemple un format texte, un format html, ou encore tout document dont les formats sont connus par leur abréviation ou leur nom commercial ou leur extension parmi lesquelles on trouve notamment :.doc et .docx, xls, rtf, ppt, xls, pdf ou open office.

L'étape d'identification du document peut être précédée ou suivie d'une étape d'importation dudit document numérique. L'importation du document numérique ou d'un ensemble de documents contenus dans dossier/répertoire peut se faire également au même moment que son identification.

Les données de forme du document numérique peuvent être déterminées par le procédé de l'invention lors de l'étape d'importation.

Le procédé permet donc d'importer au moins un document numérique et de le stocker dans un espace mémoire, par exemple la mémoire d'un composant d'un ordinateur ou d'un serveur de données.

Le stockage du document peut être effectué dans un répertoire d'un système d'exploitation d'un ordinateur.

L'importation peut être réalisée par tout moyen informatique permettant de sauvegarder les données contenues dans le document numérique. Par exemple, l'importation peut être effectuée en copiant le fichier, en utilisant une fonction « copier/coller » d'un éditeur ou encore par téléchargement du document provenant d'un autre ordinateur. L'importation peut-être également réalisée en affichant une partie ou la totalité du contenu dudit document numérique stocké sur un serveur dans un navigateur d'un ordinateur local.

Le procédé de l'invention comporte une étape de sélection, notée E_SEL, d'une base de fragments de phrases indicateurs également noté FPI signifiant « Fragment de Phrases Indicateurs ». Cette base de fragments de phrases indicateurs comprend un ensemble de TAG linguistiques, notés TAG_LIN, prédéfinis. Les TAG linguistiques peuvent comprendre des termes ou des expressions, c'est-à-dire un ensemble de termes ayant un sens pris ensemble. Cette base de FPI peut être liée à un dictionnaire morphologique qui va permettre toutes les dérivations des termes répertoriés dans cette base.

D'une manière générale, on note dans la suite de la description un TAG comme étant un terme ou un ensemble de termes formant une expression et ayant un sens syntaxique ou grammaticale.

Chaque TAG linguistique du FPI comprend une première attribution d'une valeur numérique choisie dans un premier intervalle, noté I1. Le premier intervalle est défini par une première valeur minimale, notée TAG_LIN_MIN et une première valeur maximale notée TAG_LIN_MAX.

Un dictionnaire linguistique peut être associé à la base de fragments de phrases indicateurs pour une langue donnée. Il peut exister une pluralité de dictionnaires linguistiques qui peuvent être sélectionnés dans le procédé de l'invention.

En outre, un dictionnaire morphologique comprend des données permettant de reconnaitre un TAG linguistique dit « racine » ou une expression comprenant une pluralité de termes également appelée « racine » permettant d'associer des variantes de TAG ou d'expression en fonction de règles grammaticales ou de conjugaisons. Ces données

permettent de regrouper, sous une même racine, une famille de TAG et/ou d'expressions.

Un avantage du dictionnaire morphologique de l'invention est qu'il est optimisé de sorte à permettre de générer des scores rapidement avec une pertinence optimisée. Notamment, le dictionnaire morphologique peut comprendre un nombre limité d'expressions ce qui permet d'alléger les opérations de reconnaissances de terminaisons comprises dans le dictionnaire morphologique. En outre, un autre avantage du dictionnaire morphologique de l'invention est de supprimer les déclinaisons de certaines conjugaisons non utiles dans le procédé de l'invention. A titre d'exemple, les modes impératifs, les conjugaisons de la seconde personne du singulier ainsi que les conjugaisons de la seconde personne du pluriel ne sont pas présentes dans le dictionnaire morphologique. Ce dictionnaire morphologique est spécialement adapté au procédé de l'invention de sorte à optimiser la pertinence des résultats et les temps de calculs.

Une base de fragments de phrases indicateurs comprend un ensemble de TAG linguistique, chacun ayant une valeur attribuée représentant un degré d'importance linguistique prédéfini vis-à-vis du sens d'une phrase. A titre d'exemple, l'expression « en conclusion » revêt une importance quant à ce qui va être annoncé juste après dans la phrase. D'autres exemples peuvent être cités comme : « un point important » ou encore « il est primordial » qui sont des expressions comportant une valeur attribuée proche de la limite maximale du premier intervalle.

En conséquence, la base de fragments de phrases indicateurs comprend une première attribution, notée ATT1, de valeurs à chaque TAG de la base qui représente une « importance » vis-à-vis du sens des termes qui sont supposés être exposés précédemment ou successivement à un TAG linguistique donné.

Les valeurs de la première attribution sont comprises dans un premier intervalle de valeurs. Le premier intervalle est défini par une valeur minimale et une valeur maximale.

Les valeurs sont préférentiellement prédéfinies et attribuées manuellement par un opérateur. En outre, elles peuvent être générées automatiquement selon le type de base de FPI qui a été sélectionné.

Dans un exemple simplifié de l'invention, tous les termes d'un ensemble de TAG_LIN peuvent comprendre la même valeur attribuée, notée $V1_{moy}$.

L'étape de sélection du procédé de l'invention peut comprendre également la sélection d'un thésaurus noté THE, cette étape est réalisée dans l'étape E_SEL.

Un thésaurus définit un fichier comprenant une liste de TAG sémantiques, les TAG étant notés TAG_SEM et représentent un champ lexical d'un domaine prédéfini. Le procédé de l'invention peut comprendre la sélection d'une pluralité de thésaurus par un utilisateur.

Chacun des TAG sémantiques comprend une seconde attribution, notée ATT2, de valeurs comprises dans un second intervalle, noté I2, définit par une seconde valeur minimale, notée TAG_SEM_MIN et une seconde valeur maximale TAG_SEM_MAX.

Dans un exemple simplifié de l'invention, tous les termes d'un thésaurus peuvent comprendre la même valeur attribuée, notée $V2_{moy}$.

L'étape de sélection du procédé de l'invention peut comprendre également la sélection d'un ensemble de TAG définis par un utilisateur définissant des « TAG utilisateurs », notés TAG_UTI. Les TAG utilisateurs peuvent comprendre des expressions sémantiques et/ou des termes simples.

Chaque TAG utilisateur comprend une troisième attribution, notée ATT3 de valeurs comprises dans un troisième intervalle, noté I3, défini par une troisième valeur minimale (TAG_UTI_MIN) et une troisième valeur maximale (TAG_UTI_MAX).

Dans un exemple simplifié de l'invention, tous les termes d'un ensemble de TAG utilisateurs peuvent comprendre la même valeur attribuée, notée $V3_{moy}$.

La base de fragments de phrases indicateurs peut être définie dans un fichier texte ou une base de données ou tout autre fichier numérique dont la consultation et les opérations sont autorisées. Il en est de même pour les thésaurus et les ensembles de TAG utilisateurs.

Une interface permet à un utilisateur d'éditer un fichier de TAG utilisateurs ou de sélectionner par exemple dans un menu déroulant un thésaurus. La sélection d'une langue, par exemple à partir d'une case à cocher numérique permet de définir et d'associer le thésaurus associé.

Le procédé de l'invention comprend une étape de segmentation, notée E_SEG, du premier document numérique permettant de déterminer un premier ensemble de phrases, noté P1, du premier document numérique. Lors de la reconnaissance de chacune des phrases du document numérique, les phrases sont numérotées et définissent une première séquence.

L'étape de segmentation comprend donc une identification des phrases par exemple à partir d'un analyseur syntaxique qui reconnait chaque couple {point de ponctuation – majuscule} dans le document numérique.

Dans un mode de réalisation, une partie des phrases du document numérique peuvent être identifiées ce qui permet d'appliquer le procédé de l'invention à une partie seulement d'un document numérique. Par exemple, il est possible de limiter la segmentation à un chapitre d'un document numérique, le chapitre étant délimité par des symboles ou une police ou un titre permettant de définir la partie du document auquel s'applique le procédé. L'utilisateur peut disposer de moyens de sélection d'une partie d'un texte, par exemple par une sélection à partir d'un curseur et d'une souris sur un document numérique affiché dans un afficheur.

Un avantage de pouvoir paramétrer la partie du document numérique auquel s'applique le procédé est de pré-segmenter un texte de plusieurs chapitres par exemple qui traite chacun de sujet dans des domaines différents.

Si le procédé de génération d'un résumé numérique est localement appliqué à une partie d'un document, tel qu'un chapitre par exemple, cela permet d'appliquer le procédé à différents chapitres et de générer une pluralité de résumés numériques dont le contenu peut être plus pertinent et plus proche du sens original du document numérique.

Le procédé de l'invention peut donc comporter une étape de pré-segmentation permettant d'identifier des parties d'un document et une étape de segmentation permettant d'identifier toute ou une partie des phrases du

document. Ce cas est notamment avantageux lorsque des chapitres d'un document numérique traite de sujets très différents.

Le procédé de l'invention permet, en outre, d'ordonner les phrases identifiées, les dites phrases définissant ainsi une séquence. Dans un mode de réalisation préféré, l'ordre d'apparition des phrases dans le premier document numérique est l'ordre de la séquence des phrases lors de l'étape de segmentation. Dans un mode de réalisation simple, les phrases sont simplement numérotées de la première à la dernière phrase du document numérique ou d'une partie du document numérique.

Le procédé de l'invention comprend une étape de comparaison, notée E_COM, entre les termes de chaque phrase du premier document segmenté et des TAG linguistiques de la base de fragments de phrases indicateurs et éventuellement des déclinaisons obtenues à partir d'un dictionnaire morphologique. Cette étape de comparaison permet de repérer la présence des TAG linguistiques et de leurs déclinaisons dans les phrases du texte d'origine.

Dans une alternative du procédé de l'invention, il est possible d'effectuer cette étape de comparaison sur une partie ou la totalité du document numérique et de réaliser l'étape de segmentation ensuite.

Dans un mode de réalisation amélioré du procédé de l'invention, il est possible pour chacune des phrases du texte segmenté à partir :

- d'une ou plusieurs bases de fragments de phrases indicateurs comprenant un premier ensemble de TAG linguistiques, TAG_LIN et de leurs déclinaisons ;
- d'un ou plusieurs thésaurus comprenant un second ensemble de TAG sémantiques, TAG_SEM, et ;
- d'un ensemble de TAG utilisateurs, TAG_UTI,

de comparer les termes ou expressions de ces dernières phrases avec le premier et/ou le second et/ou le troisième ensemble de TAG définis précédemment.

Dans la description qui suit et dans la définition de l'invention, nous entendons par « TAG linguistiques », les « TAG linguistiques » définis dans la base de fragments de phrases indicateurs ainsi que leurs déclinaisons déduites d'un dictionnaire morphologique lorsqu'il est utilisé.

Le procédé de l'invention comprend au moins la sélection d'une première base de fragments de phrases indicateurs définissant un premier ensemble de TAG. De manière à améliorer la cohérence des phrases identifiées selon le procédé de l'invention, un thésaurus et un ensemble de mots clefs utilisateurs peuvent être utilisés.

Le procédé de l'invention permet de lister tous les termes ou expressions de chaque phrase présente dans les trois ensembles de TAG définis précédemment.

Le procédé de l'invention comprend une étape de pondération de chaque phrase. L'étape de pondération d'une phrase comprend la sommation des valeurs attribuées de chaque TAG présent dans ladite phrase, les TAG pouvant provenir d'un des trois ensembles de TAG définis précédemment.

Une pondération permet donc de quantifier la représentativité de la phrase vis-à-vis d'au moins un FPI lié au dictionnaire morphologique, d'au moins un thésaurus ou au moins un ensemble de mots clefs sélectionnés pour le premier document numérique.

Ainsi le procédé de l'invention comprend une étape de segmentation qui permet de générer une liste de phrases ordonnées et comprenant un score obtenue par l'étape de pondération.

Dans un exemple de réalisation, un fichier constituant une base de fragments de phrases indicateurs de mots et d'expressions définissants un premier ensemble de $\{TAG_LINi\}_{i\in[1\ ;\ N]}$ est associé au document numérique.

Toujours dans cet exemple, un fichier est sélectionné représentant un thésaurus d'un domaine choisi par un utilisateur comprenant un second ensemble de TAG sémantiques $\{TAG_SEMi\}_{i\in[1\ ;\ P]}$ d'un champ lexical de ce domaine

Un opérateur défini manuellement un troisième ensemble de $\{TAG_UTIi\}_{i\epsilon[1\ ;\ K]}$ utilisateurs qu'il souhaite associer à ce document numérique.

Dans cet exemple, les trois listes de TAG $\{TAG_LINi\}_{i\epsilon[1\ ;\ N]}$, $\{TAG_SEMi\}_{i\epsilon[1\ ;\ P]}$, $\{TAG_UTIi\}_{i\epsilon[1\ ;\ K]}$ permettent de calculer les valeurs attribuées à chacun des termes de chacune des phrases identifiées dans le document numérique.

La première liste $\{TAG_LINi\}_{i\epsilon[1\ ;\ N]}$ permet notamment de repérer dans le document numérique des expressions contextualisant des phrases importantes, telles que : « en conclusion », « pour finir », « retenons que », « il est primordial que », etc. Cette liste est non représentative de tous les exemples possibles mais permet de définir un exemple précis de réalisation.

Chacune de ces expressions ou de ces termes a une valeur définie dans un premier intervalle qui peut être attribuée à chaque terme.

Si le premier intervalle est de 1 à 100. Les expressions « en conclusion », « pour finir » peuvent avoir une valeur de 70 et les expressions « retenons que », « il est primordial que » peuvent avoir une valeur de 90.

L'étape de pondération permet d'attribuer à chaque phrase du document numérique une valeur de pondération qui est par exemple la somme des valeurs de chaque terme ou expression de la phrase étant identifiés dans un des ensembles de TAG. Par exemple si une phrase comprend les deux expressions : « Pour finir, retenons que… », une valeur de la phrase peut déjà être de 70 + 90 = 160. Cette somme est, pour l'instant, calculée sans compter des valeurs potentiellement attribuées à d'autres termes de la phrase présents dans les autres listes de TAG.

Si le thésaurus « Economie » est sélectionné, des termes comme « bilan », « business plan », « entreprise », « faillite », etc peuvent définir un champ lexical que l'on souhaite appliqué dans l'extraction de phrases pertinentes d'un document. Dans cet exemple, le second intervalle est défini par une valeur minimale de 0 et une valeur maximale de 50. Dans un exemple simplifié tous les termes du thésaurus ont une valeur de 25.

En reprenant l'exemple précédent, une phrase commençant par « Pour finir, retenons que la faillite de l'entreprise A… » cumule les valeurs

de 70, 90, 25 et 25 et le score pour l'instant attribué à la phrase est de 70 + 90 + 25 + 25 = 210.

Si l'utilisateur a défini une liste de mot clef définissant des TAG_UTI tels que « 2011 » ou « camembert ». Dans cet exemple, le troisième intervalle est défini par une valeur minimale de 0 et une valeur maximale de 50. Dans un exemple simplifié tous les termes des TAG utilisateurs ont une valeur de 25.

Dans, l'exemple précédent, une phrase commençant par « Pour finir, retenons que la faillite de l'entreprise A spécialisée dans les téléviseurs est due à son étonnant changement d'activité, notamment dans le camembert en 2011. » cumule les valeurs de 70, 90, 25, 25, 25 et 25 et le score attribué à cette phrase est de 70 + 90 + 25 + 25 + 25 + 25 = 260.

Le procédé comprend une étape d'identification, notée E_IDE, d'un second ensemble de phrases, noté P2 compris dans le premier ensemble de phrases P1 formant le document numérique ayant un score supérieure à un premier seuil.

L'étape d'identification comprend la comparaison de chaque pondération de chaque phrase à une valeur définissant un seuil prédéfini. Le seuil prédéfini peut être fixé à l'avance ou modifié à tout moment au moyen d'une interface.

Le procédé de l'invention comprend en outre une étape de paramétrage du procédé de l'invention définie ci-après.

L'étape d'identification permet la génération d'une seconde liste de phrases dont le score est supérieur à un seuil prédéfini. Dans une alternative il est possible de définir un nombre maximum de phrases du résumé numérique qu'un utilisateur souhaite définir. Ce nombre maximum de phrases peut s'exprimer en fonction d'un pourcentage du nombre de phrases du document ou de la partie du document auquel s'applique le procédé de l'invention. Les phrases ayant le meilleurs score soit au-dessus d'un seuil soit déterminé par un nombre de phrase maximum définissent un second ensemble de phrases P2.

Les phrases de la seconde liste sont ordonnées et comprennent une numérotation, par exemple la même numérotation que dans la première liste.

Ainsi si la première liste comprend par exemple 100 phrases numérotées de 1 à 100 et que seules 5 phrases ont été retenues dans la seconde liste, dont les phrases numérotées 20, 30, 40, 50 et 61, leur numérotation peut être préservées dans la seconde liste.

Le procédé sera toujours capable de les ordonner par exemple pour les afficher dans un ordre précis en comparant les numérotations de chacune des phrases. Il sera tout aussi simple d'établir la comparaison suivante : 20 < 30 < 40 < 50 < 61, pour établir un ordre que de renuméroter les phrases sélectionnées suite à l'étape de comparaison de leur score avec un seuil prédéfini.

Un avantage de la seconde liste de TAG est qu'elle permet d'orienter l'identification des phrases du document numérique selon un thesaurus formé par un ensemble de TAG représentatif d'un domaine précis.

Ainsi il peut être généré autant de résumés numériques du premier document numérique que de fichiers différents parmi lesquels on trouve par exemple le FPI, un fichier de langues, un thésaurus particulier ou un fichier comprenant une liste de TAG utilisateur.

L'invention permet de configurer un rapport entre les intervalles I1, I2 et I3 ou de leur données représentatives telle que la valeur moyenne des valeurs attribuées d'un intervalle ou le centre de chaque intervalle.

Une première configuration consiste à choisir un intervalle I2 inclus dans l'intervalle I1. De manière analogue, un intervalle I3 peut être choisi de sorte à être inclus dans l'intervalle I1. C'est-à-dire que la borne supérieure du premier intervalle I1 est supérieure à la borne supérieure du second intervalle I2. Identiquement, la borne supérieure du premier intervalle I1 peut être également supérieure à la borne supérieure du troisième intervalle I3.

Ces configurations sont particulièrement avantageuses dans la mesure où de nombreux tests ont été conduits permettant d'aboutir à des

résultats pertinents de résumés générés avec cette configuration. Etant donné que l'intervalle I1 représente des valeurs d'un ensemble de FPI défini manuellement conjointement à un dictionnaire morphologique, cet ajustement a été défini selon une analyse d'un grand nombre de résultats et d'essais. En effet, les FPI ont été définis à partir de la collecte et l'analyse de fragments de phrases associés à une importance du sens des phrases comprenant ces FPI. On comprend alors que l'ajustement des intervalles requiert une importance lors de la configuration.

En effet, un résumé pertinent ne peut être jugé qu'en comparaison de la lecture du texte original dont il est issu. A cet effet, de nombreux essais ont permis de définir des intervalles I1, I2 et I3 et de leur relations permettant de générer les phrases ayant les meilleurs scores reflétant le mieux la nature du texte dont le résumé est généré.

Une configuration particulièrement avantageuse pour optimiser la cohérence et la fidélité du document numérique dans l'identification des phrases du procédé peut être définie. Notamment, la définition de la borne maximale du premier intervalle peut être prise sensiblement égale à la moitié de la borne maximale du second ou du troisième intervalle. Cette configuration permet de privilégier les formes syntaxiques d'un document représentant des propos ayant une importance quant au sens.

Avantageusement, ce paramétrage peut être configuré selon la nature des documents dont le procédé réalise l'identification des phrases. Par exemple, des documents de brevets, des publications scientifiques, des brochures commerciales, des manuels, des guides, des modes d'emploi, des livres tels que des romans comporte chacun un lexique morphologique propre à la nature du document. En conséquence les données caractéristiques des intervalles I1, I2 et I3 peuvent être adaptées au cas par cas.

Le procédé de l'invention comprend dans un mode amélioré, une étape préliminaire de paramétrage au moyen d'une interface permettant à un opérateur d'adapter à ses besoins l'application du procédé au texte numérique.

Un premier paramétrage comprend la définition d'une première valeur représentant le degré de condensation du document numérique. Cette

valeur représente un rapport entre le nombre de phrases identifiées par le procédé de l'invention et le nombre de phrases du document numérique ou d'une partie identifiée de ce dernier.

On entend par meilleur score : le score le plus élevé d'une phrase lorsque les valeurs attribuées sont additionnées positivement ou encore les scores dépassant un certain seuil prédéfini.

L'utilisateur peut par exemple choir d'afficher les phrases identifiées ayant le meilleur score et représentant 10% du nombre de phrases du document. En conséquence, le procédé de l'invention choisira sur 100 phrases d'un document numérique, les 10 phrases ayant le meilleur score.

On appelle « taux de condensation », le rapport entre le nombre de données générées dans le résumé numérique et le nombre de données du document numérique. Les données peuvent être exprimées en nombre de caractères, en nombre de mots, en nombre de phrases, en nombre de paragraphes ou encore en nombre de page selon les différents modes de réalisation de l'invention.

Le procédé de l'invention concerne un procédé d'identification de phrases d'un document numérique qui peuvent être générées selon une symbologie particulière dans leur contexte initial. Le contexte initial est défini par l'affichage d'une phrase parmi les autres phrases du document numérique, c'est-à-dire normalement lorsque que le texte du document est simplement affiché.

La symbologie particulière peut concerner une couleur, une police ou une taille de police. Ainsi lorsque le procédé s'applique par exemple à un texte numérique affiché dans un navigateur internet, les phrases identifiées selon le procédé de l'invention peuvent apparaitre en gras avec un corps de police supérieur au corps de police des phrases non identifiées. D'autres possibilités de démarcation facilitant la lecture dite « en diagonale » d'un texte peuvent être combinées ensembles. La génération des phrases identifiées selon le procédé de l'invention avec une symbologie particulière pour être reconnaissables, lorsqu'elles sont générées dans leur contexte initial, peuvent l'être dans tout afficheur ou tout logiciel d'affichage numérique tel qu'un éditeur ou un navigateur numérique.

L'invention permet de générer les phrases identifiées dans la même police de caractère mais avec une variation des formats correspondants aux scores calculés pour chacune des phrases. Par exemple, les phrases de score plus conséquent se verront attribuer un affichage plus grand. Les phrases de score moins conséquent se verront attribuer un affichage plus petit. Un dégradé de cet affichage est appliqué à tout le document source. Les phrases pouvant véhiculer des informations importantes sont affichées en grands caractères. Inversement, celles de moindre importance sont affichées en petits caractères. Une échelle de grandeur de cet affichage permet à l'utilisateur de parcourir en un seul coup d'œil le document et/ou son résumé .

Le procédé peut s'appliquer à un corpus de N documents numériques, par exemple, par la génération d'un résumé numérique de l'ensemble des phrases de tous les documents numériques. Il est également possible de spécifier un taux de condensation pour chacun des documents. Le procédé exécute alors le procédé de l'invention sur une liste de documents et permet d'afficher alors une synthèse numérique. La synthèse numérique est la juxtaposition d'une pluralité de résumés numériques générés par le procédé de l'invention appliqué à plusieurs documents numériques.

La synthèse numérique est générée par le procédé de l'invention auquel deux étapes supplémentaires ont été ajoutées. Il y a alors une première étape de paramétrage permettant de spécifier le taux de condensation de chaque résumé du numérique contribuant à l'élaboration de la synthèse numérique. Il y a une étape de création de la synthèse par la juxtaposition d'une pluralité de résumés numériques.

Prenons par exemple trois documents numériques D1, D2, D3 dont le procédé est exécuté pour générer une synthèse numérique. Le procédé de l'invention s'applique à chacun des documents numériques en spécifiant dans le paramétrage d'une interface le taux de condensation de chacun des résumés de chacun des documents.

Par exemple, un premier résumé R1 comprend un taux de condensation de 20% de D1, un second résumé R2 comprend un taux de condensation de 10% de D2, un troisième résumé comprend un taux de

condensation de 5% de D1. La synthèse numérique S1 comprend alors la juxtaposition des trois résumés R1, R2 et R3.

L'invention comprend un dispositif de génération d'au moins un résumé numérique. Ce dernier comprend des moyens de calculs permettant de mettre en œuvre les étapes du procédé, un afficheur permettant d'afficher le document numérique et/ou le résumé numérique. En outre, le dispositif de l'invention comprend des moyens de sélection de paramètres de la configuration ou du paramétrage du procédé.

En outre, l'afficheur peut comprendre un navigateur disposant :

- d'une première fenêtre permettant d'afficher d'une part une pluralité de symboles représentant des documents ordonnés selon une séquence donnée et d'autre part les titres ou des références des documents de manière à les rendre identifiables ;
- d'une seconde fenêtre permettant d'afficher les résumés de chacun des documents, le résumé étant généré au moyen du procédé de l'invention.

Dans la seconde fenêtre l'ordre d'affichage des résumés, par exemple les uns au-dessous des autres, peut être fidèle à la séquence d'affichage des documents. Ainsi, pour un utilisateur il y a une cohérence entre l'ordre d'affichage des documents ou de leurs symboles dans une première fenêtre et les résumés qui sont dans une seconde fenêtre préférentiellement disposée à côté de la première fenêtre.

Dans un mode de réalisation, un symbole est généré à proximité de chaque phrase du résumé numérique. Chaque symbole est activable par des moyens de sélection contrôlés par un utilisateur tels qu'une souris et curseur ou un toucher tactile sur un écran tactile.

Le symbole peut être un ou plusieurs caractère(s) alphanumérique(s), par exemple tels que des signes « + » ou « - ». Chaque symbole peut être généré à proximité de chacune des phrases du résumé numérique. Les symboles peuvent tous être générés dans une même partie, par exemple à gauche ou droite du résumé affiché sur la même ligne que le

début ou la fin d'une phrase. Ils peuvent également être affichés dans le texte du résumé numérique après chaque point ou majuscule du texte.

L'activation de ces signes permet de générer l'affichage des phrases consécutives ou précédentes la phase positionnée à côté du signe. Cette caractéristique permet de contextualiser une phrase qui aurait perdu du sens lors de son extraction du document numérique.

Par ailleurs, un double-clic sur une phrase du résumé généré permet sa suppression de la liste des phrases retenues pour le cas où l'utilisateur ne souhaiterait pas disposer de cette phrase dans le résumé final.

Ainsi le dispositif de l'invention, permet d'offrir un moyen simple à l'utilisateur de récupérer un degré de cohérence et de fidélité du résumé numérique vis-à-vis du document numérique par une action simple et rapide.

Une activation du signe permet d'afficher immédiatement la phrase précédente et/ou celle suivante la phrase associée à un symbole activé. Un double-clic sur la phrase permet sa suppression de l'affichage.

Selon le paramétrage effectué, une action sur un signe permet d'afficher une ou une pluralité de phrases avant ou après la phrase dont on souhaite éclairer le contexte. Cette donnée est paramétrable dans un mode de réalisation.

Enfin l'invention comprend de nombreux avantages. La définition des TAG_LIN de la base de fragments de phrases indicateurs permet au procédé de prendre en compte des expressions et des termes qui représentent une forme d'importance dans l'extraction des points, c'est-à-dire des phrases, importants d'un document qui dépendent de la structure morphologique d'une langue donnée.

Le thésaurus permet d'orienter la génération d'un résumé selon un axe sémantique particulier, par exemple le secteur de l'automobile. Enfin, les mots clefs utilisateurs permettent de prendre en compte des considérations de recherches spécifiques d'un individu.

Ainsi, chaque résumé numérique selon les critères de sélection de fichiers et/ou de définition de TAG permet de générer un résumé « sur mesure ». Ce dernier est généré avec une fidélité et une cohérence vis-à-vis du document numérique qui peuvent être corrigées ou contextualisées.

REVENDICATIONS

1. Procédé d'identification d'un ensemble de phrases d'un premier document numérique (D1), caractérisé en ce qu'il comprend :
   - une étape d'importation (E_IMP) du premier document numérique (D1) dans au moins un format prédéfini permettant : soit d'afficher le document dans une première interface soit de le stocker dans une mémoire ;
   - une étape de sélection (E_SEL) d'une base de fragments de phrases indicateurs (FPI) comprenant un ensemble de TAG linguistiques (TAG_LIN), chacun des TAG linguistiques comprenant une première attribution de valeurs numériques choisies dans un premier intervalle (I1) défini par une première valeur minimale (TAG_LIN_MIN) et une première valeur maximale (TAG_LIN_MAX) ;
   - l'étape de sélection comprenant également la sélection d'un thésaurus (THE) définissant un fichier comprenant une liste de TAG sémantiques (TAG_SEM) d'un domaine, chacun des TAG sémantiques comprenant une seconde attribution (ATT2) de valeurs pour chaque TAG sémantique comprises dans un second intervalle (I2) défini par une seconde valeur minimale (TAG_SEM_MIN) et une seconde valeur maximale (TAG_SEM_MAX), la seconde valeur maximale (TAG_SEM_MAX) étant inférieure à la première valeur maximale (TAG_LIN_MAX) du premier intervalle (I1) ;
   - une étape de segmentation (E_SEG) du premier document numérique permettant de :
     - déterminer un premier ensemble de phrases (P1) du premier document (D1) ;
     - numéroter les phrases de ce premier ensemble définissant une première séquence ;
   - une étape de comparaison (E_COM) des termes de chaque phrase du premier document segmenté et des TAG

linguistiques de la base de fragments de phrases indicateurs permettant de repérer la présence des TAG linguistiques dans lesdites phrases ;

- une étape de pondération (E_PON) de chacune des phrases par attribution d'un premier score correspondant à la somme des valeurs de chaque TAG linguistique repérés dans chacune des phrases ;
- l'étape de pondération (E_PON) de chacune des phrases comprenant en outre une attribution d'un second score correspondant à la somme des valeurs de chaque TAG sémantiques repérés dans chacune des phrases.
- une étape d'identification (E_IDE) d'un second ensemble de phrases (P2) compris dans le premier ensemble de phrases,
    - le premier score ou ;
    - le second score ou ;
    - la somme du premier et du second score,

  des phrases du second ensemble étant supérieur à un premier seuil.

2. Procédé d'identification d'un ensemble de phrases d'un document numérique selon la revendication 1, caractérisé en ce que le premier seuil est calculé à partir d'un taux de condensation défini par le nombre de phrases souhaité par un utilisateur du second ensemble sur le nombre total de phrases du premier ensemble de phrases.

3. Procédé d'identification d'un ensemble de phrases d'un document numérique selon la revendication 1, caractérisé en ce que le premier seuil est calculé à partir d'un taux de condensation défini par le nombre de termes souhaités par un utilisateur du second ensemble de phrases sur le nombre total de termes du premier ensemble de phrases.

4. Procédé d'identification d'un ensemble de phrases d'un document numérique selon la revendication 1, caractérisé en ce qu'une interface permette de configurer le taux de condensation.

5. Procédé d'identification d'un ensemble de phrases d'un premier document numérique selon l'une quelconque des revendications 1 à 4, caractérisé en ce qu'une étape d'affichage au moyen d'une interface du premier document numérique comprend la génération des phrases identifiées selon une taille de caractère plus importante que les phrases non identifiées.

6. Procédé d'identification d'un ensemble de phrases d'un premier document numérique selon l'une quelconque des revendications 1 à 5, caractérisé en ce que l'étape de comparaison (E_COM) comprend la détermination de termes racines des TAG linguistiques du FPI à partir d'un dictionnaire morphologique et la comparaison des déclinaisons des termes racines des TAG linguistiques avec chaque phrase du document numérique.

7. Procédé d'identification d'un ensemble de phrases d'un premier document numérique selon l'une quelconque des revendications 1 à 6, caractérisé en ce que :
   - l'étape de sélection (E_SEL) comprend la sélection d'un ensemble de TAG définis par un utilisateur définissant des TAG utilisateurs (TAG_UTI) comprend des expressions sémantiques et/ou des termes, chacun des TAG utilisateurs comprenant une troisième attribution (ATT3) de valeurs pour chaque TAG utilisateurs comprises dans un troisième intervalle (I3) définit une troisième valeur minimale (TAG_UTI_MIN) et une troisième valeur maximale (TAG_UTI_MAX) ;
   - l'étape de pondération (E_PON) de chacune des phrases par attribution d'un troisième score correspondant à la somme des valeurs de chaque TAG utilisateurs repérés dans chacune des phrases.

8. Procédé d'identification d'un ensemble de phrases d'un premier document numérique selon l'une quelconque des revendications 1 à 7, caractérisé en ce que l'étape de pondération comprend la somme des premiers, seconds et/ou troisièmes score pour chacune des

phrases du document numérique, définissant ainsi un poids sémantique, le poids sémantique de chaque phrase étant comparé à un seuil prédéfini dans l'étape d'identification.

9. Procédé d'identification d'un ensemble de phrases d'un premier document numérique selon l'une quelconque des revendications 1 à 8, caractérisé en ce que la valeur moyenne des valeurs de la seconde attribution (ATT2) est dans un intervalle représentant 20% du premier l'intervalle (I1) centré sur la valeur moyenne des valeurs de la première attribution.

10. Procédé d'identification d'un ensemble de phrases d'un premier document numérique selon l'une quelconque des revendications 1 à 8, caractérisé en ce que la valeur moyenne des valeurs de la troisième attribution (ATT3) est dans un intervalle représentant 20% du premier l'intervalle (I1) centré sur la valeur moyenne des valeurs de la première attribution.

11. Procédé de génération (E_GEN) d'un document numérique, noté « résumé numérique », comprenant la génération et l'affichage sur un afficheur du second ensemble de phrases, les dites phrases étant identifiées à partir du procédé d'identification de l'une quelconque des revendications 1 à 10, selon une séquence ordonnée par une numérotation croissante.

12. Procédé de génération d'un document numérique selon la revendication 11, caractérisé en ce que le résumé numérique généré comprend des symboles activables, un symbole activable étant associé à chacune des phrases du second ensemble, les phrases du résumé numérique et les symboles activables étant affichées sur un afficheur de manière à ce que les symboles activables soient affichées à proximité des phrases, l'activation d'au moins un symbole activable d'une phrase sélectionnée générant un second résumé numérique, le second résumé numérique comportant des phrases ordonnées dont la numérotation est successive, cet ensemble comportant ladite phrase sélectionnée et un premier ensemble de phrases dont la numérotation

précède celle de la phrase sélectionnée et un second ensemble de phrases dont la numérotation succède à celle de la phrase sélectionnée.

13. Procédé de génération d'un document numérique selon la revendication 12, caractérisé en ce que l'activation d'un symbole activable est réalisé au moyen d'un clic de souris pour ordinateur ou d'un survol d'un curseur sur des données activables ou d'un toucher tactile dans une zone comprenant le symbole activable.

14. Procédé de génération d'un document numérique selon la revendication 12, caractérisé en ce que le symbole activable est un caractère alphanumérique.

15. Procédé de génération d'un document numérique selon la revendication 12, caractérisé en ce que le symbole activable est un numéro représentant le numéro de la phrase dans le premier document.

16. Procédé de génération d'un document numérique, appelé « synthèse numérique », caractérisé en ce le procédé selon l'une quelconque des revendications 11 à 15 est appliqué à un ensemble de documents numériques de manière à générer une pluralité de résumés numériques, ledit procédé comportant une étape de génération d'une synthèse numérique à partir de la définition d'un paramètre, dit de taux de répartition, représentant la quantification des données de chaque résumé numérique présente dans la synthèse et d'un second taux de condensation de chaque résumé numérique, la synthèse numérique comprenant un ensemble de phrases ordonnées et sélectionnées en fonction du taux de répartition et du second taux de condensation de chacun des résumé numérique.

17. Dispositif de génération d'un document numérique comportant un afficheur permettant d'afficher au moins un document numérique, un calculateur permettant de mettre en œuvre les étapes du procédé de l'une des revendications précédentes, une interface permettant de

paramétrer au moins un premier taux de condensation, un système de commandes permettant de lancer la génération d'un premier résumé du numérique.

18. Dispositif de génération d'un document numérique selon la revendication 17, caractérisé en ce que le système de commandes permet de lancer la génération d'un second résumé numérique du premier résumé numérique.

19. Dispositif de génération d'un document numérique selon la revendication 17, caractérisé en ce que l'interface comprend une première fenêtre permettant d'afficher un ensemble de documents numériques et une seconde fenêtre permettant d'afficher un ensemble de résumés numériques correspondant au résumé de chaque document de la première fenêtre.

20. Dispositif de génération d'un document numérique selon la revendication 17, caractérisé en ce que l'interface comprend des premiers moyens de sélection d'un taux de condensation d'un résumé numérique, des seconds moyens de sélection d'un thésaurus parmi, une liste de thésaurus prédéfinie et des moyens permettant de définir des TAG d'un utilisateur.

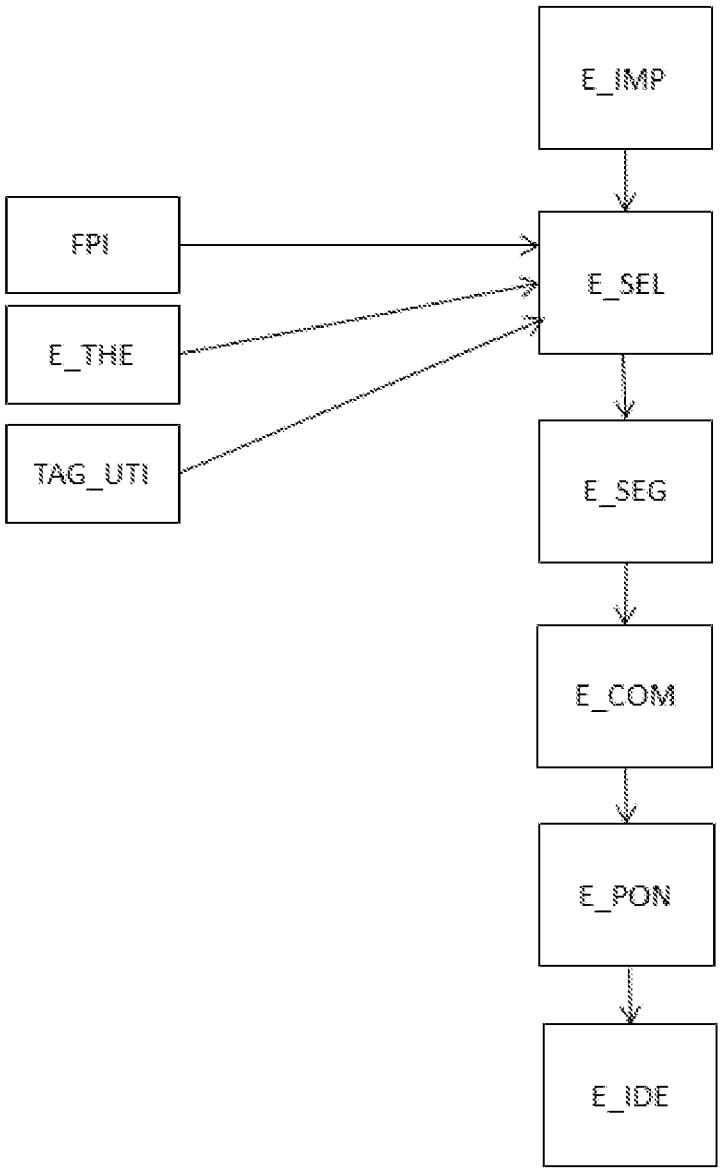


FIG.1

INTERNATIONAL SEARCH REPORT

International application No
PCT/FR2013/050269

A. CLASSIFICATION OF SUBJECT MATTER
INV. 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, INSPEC

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	Abderrafih Lehmam: "Essential Summarizer: innovative automatic text summarization software in twenty languages", Proceeding RIAO '10 Adaptivity, Personalization and Fusion of Heterogeneous Information, 2010, pages 216-217, XP055045639, Paris, France Retrieved from the Internet: URL:http://dl.acm.org/citation.cfm?id=1937055.1937111 [retrieved on 2012-11-26]	1-12, 16-20
Y	the whole document	12-15

-/--

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

[ ] 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
17 May 2013	28/05/2013

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	Fournier, Christophe

INTERNATIONAL SEARCH REPORT

International application No
PCT/FR2013/050269

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	ABDERRAFIH LEHMAM: "Text structuration leading to an automatic summary system: RAFI", INFORMATION PROCESSING & MANAGEMENT, vol. 35, no. 2, 1 March 1999 (1999-03-01), pages 181-191, XP055045640, ISSN: 0306-4573, DOI: 10.1016/S0306-4573(98)00043-0 abstract page 182, last paragraph - page 183, paragraph 1 page 183, paragraph 5 - paragraph 7 page 185, line 5 - page 190, line 4; figures 2,3	1-4,12, 18,19
Y	Abderrafih Lehmam: "i-expo 2010 : Mining Essential résume en direct", Archimag.com, 10 June 2010 (2010-06-10), pages 1-2, XP055045642, Retrieved from the Internet: URL:http://www.archimag.com/article/i-expo-2010-mining-essential-r%C3%A9sume-en-direct [retrieved on 2012-11-26] the whole document	12-15
A	Abderrafih Lehmam: "Le résumé automatique, face au déluge informationnel en français et en arabe", Les actes du colloque Gouvernance des institutions et Intelligence Economique, 29 June 2008 (2008-06-29), pages 1-22, XP055045634, Algers Retrieved from the Internet: URL:http://www.veille.ma/IMG/pdf/gouvernance-ie-actes/resume-automatique-abderrafih-lehmam.pdf [retrieved on 2012-11-26] the whole document	1-20

RAPPORT DE RECHERCHE INTERNATIONALE

Demande internationale n°
PCT/FR2013/050269

A. CLASSEMENT DE L'OBJET DE LA DEMANDE
INV. 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, INSPEC

C. DOCUMENTS CONSIDERES COMME PERTINENTS

Catégorie*	Identification des documents cités, avec, le cas échéant, l'indication des passages pertinents	no. des revendications visées
X	Abderrafih Lehmam: "Essential Summarizer: innovative automatic text summarization software in twenty languages", Proceeding RIAO '10 Adaptivity, Personalization and Fusion of Heterogeneous Information, 2010, pages 216-217, XP055045639, Paris, France Extrait de l'Internet: URL:http://dl.acm.org/citation.cfm?id=1937055.1937111 [extrait le 2012-11-26]	1-12, 16-20
Y	le document en entier	12-15
	----- -/--	

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

[ ] 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
17 mai 2013	28/05/2013

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	Fournier, Christophe

RAPPORT DE RECHERCHE INTERNATIONALE

Demande internationale n°
PCT/FR2013/050269

C(suite). DOCUMENTS CONSIDERES COMME PERTINENTS

Catégorie*	Identification des documents cités, avec, le cas échéant, l'indication des passages pertinents	no. des revendications visées
X	ABDERRAFIH LEHMAM: "Text structuration leading to an automatic summary system: RAFI", INFORMATION PROCESSING & MANAGEMENT, vol. 35, no. 2, 1 mars 1999 (1999-03-01), pages 181-191, XP055045640, ISSN: 0306-4573, DOI: 10.1016/S0306-4573(98)00043-0 abrégé page 182, dernier alinéa - page 183, alinéa 1 page 183, alinéa 5 - alinéa 7 page 185, ligne 5 - page 190, ligne 4; figures 2,3 -----	1-4,12, 18,19
Y	Abderrafih Lehmam: "i-expo 2010 : Mining Essential résume en direct", Archimag.com, 10 juin 2010 (2010-06-10), pages 1-2, XP055045642, Extrait de l'Internet: URL:http://www.archimag.com/article/i-expo-2010-mining-essential-r%C3%A9sume-en-direct [extrait le 2012-11-26] le document en entier -----	12-15
A	Abderrafih Lehmam: "Le résumé automatique, face au déluge informationnel en français et en arabe", Les actes du colloque Gouvernance des institutions et Intelligence Economique, 29 juin 2008 (2008-06-29), pages 1-22, XP055045634, Algers Extrait de l'Internet: URL:http://www.veille.ma/IMG/pdf/gouvernance-ie-actes/resume-automatique-abderrafih-lehmam.pdf [extrait le 2012-11-26] le document en entier -----	1-20