(19) RÉPUBLIQUE FRANÇAISE

INSTITUT NATIONAL
DE LA PROPRIÉTÉ INDUSTRIELLE

PARIS

(11) **N° de publication :** **2 694 984**
(à n'utiliser que pour les commandes de reproduction)

(21) **N° d'enregistrement national :** **92 11714**

(51) **Int Cl⁵ : G 06 F 15/40**

(12) DEMANDE DE BREVET D'INVENTION A1

(22) **Date de dépôt** : 02.10.92.

(30) **Priorité** : 20.08.92 US 932522.

(43) **Date de la mise à disposition du public de la demande :** 25.02.94 Bulletin 94/08.

(56) **Liste des documents cités dans le rapport de recherche préliminaire :** *Se reporter à la fin du présent fascicule.*

(60) **Références à d'autres documents nationaux apparentés :**

(71) **Demandeur(s) :** *Gouvernement des Etats-Unis représenté par le Directeur de NATIONAL SECURITY AGENCY* — US.

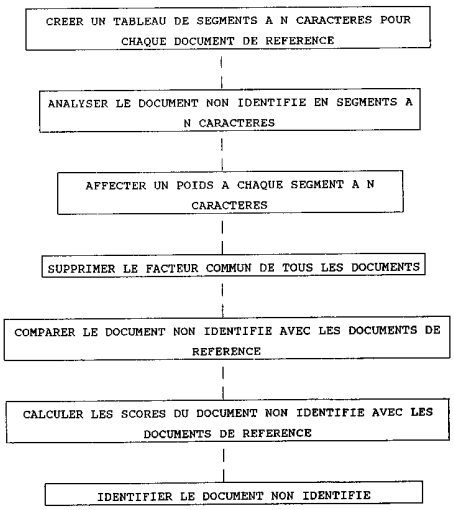
(72) **Inventeur(s) :** Damashek Marc.

(73) **Titulaire(s) :**

(74) **Mandataire :** Martinet & Lapoux.

(54) Procédé d'identification, de récupération et de classement de documents.

(57) Un procédé pour identifier, récupérer ou classer des documents par langue ou par thème inclut les étapes de créer un tableau de segments à n caractères pour chaque document d'une base de données, d'analyser un document non identifié ou question en segments à n caractères, d'affecter un poids à chaque segment à n caractères, de supprimer le facteur commun des segments à n caractères, de comparer chaque document non identifié ou question à chaque document de la base de données, de calculer un score pour le document non identifié ou question par rapport à chaque document de la base de données pour la similarité. En se basant sur le score de similarité le document ou question est identifié, récupéré ou classé par rapport à la langue ou au thème.



FR 2 694 984 - A1

Procédé d'identification, de récupération et de classement de documents

La présente invention a trait à un procédé de récupération de document, et plus particulièrement à un procédé de récupération de document capable de chercher un texte complet sans avoir besoin de mot-clé ou d'information sur le contexte. Ce procédé peut être utilisé pour identifier, récupérer et classer des documents par thème ou par langue. Ce procédé est également utile pour identifier, récupérer et classer toute forme de communication comme des signaux acoustiques (par exemple, de la parole) et des symboles graphiques (par exemple, des images) qui peuvent être représentés dans un format lisible par machine.

Dans "DARDA Eyes Smart Text Sifters", article publié de R. Colin Johnson dans "Electronic Engineering Times", 17 février 1992, page 35, il est indiqué que des efforts importants de recherche ont été accomplis pour trouver de meilleures méthodes pour chercher dans des bases de données de texte dans le but de récupérer des documents d'intérêt pour l'utilisateur. Il est indiqué que plusieurs problèmes fondamentaux demeurent dans la réalisation de percées significatives.

Une technique pour améliorer les recherches est de créer un matériel spécialisé qui peut traiter l'information plus rapidement. Le problème avec cette approche est que les améliorations dans la vitesse de traitement ne se sont pas faites au rythme auquel les informations d'une base de données ont augmenté. Il est dit qu'une avancée théorique fondamentale est

nécessaire pour améliorer la récupération des informations depuis les grandes bases de données.

Les systèmes conventionnels pour récupérer l'information sont encore fondés sur l'utilisation de mots-clés ou de parties de phrases avec des opérateurs (c'est-à-dire "et", "ou", "non") pour identifier des documents désirés. Le problème de cette technique est que des documents peuvent contenir un synonyme du mot-clé plutôt que le mot-clé lui-même (par exemple voiture pour automobile) ou une forme dérivée du mot-clé (par exemple récupération pour récupérer). De tels systèmes sont particulièrement sensibles à l'orthographe ou à des erreurs de transmission de données en entrée. Les opérateurs peuvent également être d'un emploi difficile. Des problèmes supplémentaires sont : d'identifier des mots-clés appropriés, d'identifier des synonymes appropriés, et de récupérer des documents soit insuffisants, volumineux et/ou hors sujet. Classiquement, une vaste table de synonymes est utilisée pour atténuer ces problèmes. Mais cette méthode accroît les besoins en mémoire et ralentit le temps d'exécution.

Un autre problème avec les recherches par mot-clé est que la signification du mot-clé dépend généralement du contexte dans lequel il est utilisé. Par conséquent, sans quelques indications sur le contexte désiré du mot-clé, les chances de récupérer des documents non attendus sont grandes. Des approches antérieures à la récupération de documents ont tenté de surmonter ce problème en ajoutant des informations de contexte à la recherche, en utilisant des techniques telles que des vecteurs de contexte, des graphes conceptuels, des réseaux sémantiques et des réseaux d'inférence. Ces techniques augmentent

elles aussi le besoin en mémoire et ralentissent le temps d'exécution. L'ajout d'informations de contexte est également une tâche nécessitant un temps significatif d'une personne formée.

Dans "Global Text Matching for Information Retrieval", un article publié de G. Salton et C. Buckley, dans "Science", volume 253, 30 août 1991, pages 1012 à 1015, il est indiqué que l'analyse de texte qui utilise des synonymes est volumineuse et que l'analyse de texte qui utilise une approche basée sur les connaissances est complexe. Ce même article indique que la compréhension du texte doit se fonder sur le contexte et sur la reconnaissance de parties de texte (c'est-à-dire parties de texte, paragraphes ou phrases).

Dans "Developments in Automatic Text Retrieval", un article publié de G. Salton, dans "Science", volume 253, 30 août 1991, pages 974 à 980, l'état actuel de la récupération de document est résumé. Il est indiqué que l'analyse de texte est un problème en raison de la nécessité de n'extraire que les documents intéressants depuis de grandes bases de données. La solution classique à ce problème a été de générer des identificateurs de contenu. Cela a été fait parce ce que la signification d'un mot ne peut pas être déterminé correctement en consultant un dictionnaire sans tenir compte du contexte dans lequel ce mot est employé. Il est indiqué que les mots dans le texte peuvent aussi être utilisés pour l'identification de contexte. De tels systèmes de récupération sont définis comme des systèmes de récupération de texte complet.

Dans "N-gram Statistics for Natural Language Understanding and Text processing", un article publié de C. Suen dans "IEEE Transactions On Pattern Analysis and Machine Intelligence", vol. PAMI-1, N° 2, avril 1979, deux procédés de traitement du langage naturel sont décrits, l'un utilisant des mots-clés et un dictionnaire et l'autre des segments à n caractères. Dans l'approche par mot-clé, des mots sont comparés. Dans l'approche par segment à n caractères, des chaînes de lettres sont comparées. La comparaison des chaînes de lettres est plus rapide et nécessite moins de mémoire qu'un procédé par mot-clé et dictionnaire.

Dans le brevet US-A-5 020 019 intitulé "Document Retrieval System" est décrit un système qui cherche des documents en utilisant des mots-clés avec une faculté d'apprentissage qui permet à l'utilisateur d'affecter des poids aux différents mots-clés en réponse aux résultats d'une précédente recherche. La présente invention n'utilise pas une approche par mot-clé.

Dans le brevet US US-A-4 985 863 intitulé "Document Storage and Retrieval" est décrit un procédé dans lequel les documents sont stockés par parties. Des parties de texte, plutôt que des mots-clés, sont alors utilisées pour récupérer des documents similaires. La présente invention n'utilise une approche ni par mot-clé, ni par partie.

Dans le brevet US-A-4 849 898 intitulé "Method and Apparatus to Identify the Relation of Meanings Between Words in Texte Expression" est décrit un procédé qui utilise une analyse sémantique par lettre des mots-clés et des mots d'un document

pour déterminer si ces mots signifient la même chose. Ce procédé est utilisé pour récupérer des documents ou des parties de documents qui traitent du même thème que les mots-clés. La présente invention n'utilise pas une analyse sémantique.

Dans le brevet US-A-4 823 306 intitulé "Text Search System" est décrit un procédé qui génère des synonymes de mots-clés. Différentes valeurs sont alors affectées à chaque synonyme pour guider la recherche. La présente invention ne génère pas de synonymes.

Dans le brevet US-A-4 775 956 intitulé "Method and System for Information Storing and Retrieval Using Word Stems and Derivative Pattern Codes Representing Families of Affixes" est décrit un procédé qui utilise un ensemble général d'affixes qui sont utilisés pour modifier chaque radical de mot-clé. Ce procédé réduit les besoins en mémoire qui seraient sinon nécessaires pour stocker les synonymes de chaque mot-clé. La présente invention ne modifie pas les radicaux des mots-clés.

Dans le brevet US-A-4 358 824 intitulé "Office Correspondence Storage and Retrieval System" est décrit un procédé qui réduit des documents à des résumés en enregistrant les mots-clés utilisés dans chaque document. Les mots-clés sont alors utilisés pour rechercher les documents désirés. La présente invention ne remplace pas le texte de documents mémorisés par des résumés composés de mots-clés.

La présente invention vise à rémédier aux inconvénients précités selon la technique antérieure

en fournissant un procédé de détermination de la langue dans laquelle un document donné est écrit.

A cette fin, un procédé pour déterminer la langue d'un texte donné, comprend les étapes de :

(a) analyser ledit corps de texte en une pluralité de segments à n caractères, chaque segment à n caractères comprenant le contenu de n emplacements de caractère/espace consécutifs dudit corps de texte où n a une valeur d'au moins un;

(b) comparer chacun desdits segments à n caractères avec une pluralité d'ensembles de segments à n caractères où chaque ensemble de segments à n caractères représente une langue différente; et

(c) identifier la langue dudit corps de texte comme étant écrit dans la langue d'un ensemble particulier de segments à n caractères quand le pourcentage d'appariement à l'étape (b) pour un ensemble particulier de segments à n caractères est au moins égal à la valeur prescrite pour cet ensemble de segments à n caractères et plus élevé que le pourcentage d'appariement à l'étape (b) pour les autres ensembles de segments à n caractères.

Pour identifier un document, l'invention comprend les étapes de :

(a) créer un tableau de référence de segments à n caractères pour chacun d'une pluralité de documents de référence qui sont dans un format lisible par machine, chaque segment à n caractères comprenant au moins un élément consécutif du document;

(b) analyser un document non identifié qui est dans un format lisible par machine en une pluralité de segments à n caractères, chaque segment à n caractères comprenant au moins un élément consécutif de document;

(c) affecter un poids à chaque segment à n caractères de l'étape (a);

(d) affecter un poids à chaque segment à n caractères de l'étape (b);

(e) supprimer le facteur commun parmi ladite pluralité de tableaux de référence;

(f) comparer chaque segment à n caractères de document non identifié à chaque segment à n caractères desdits tableaux de référence pour produire une liste de scores qui indiquent le degré de similarité entre ledit document non identifié et chacun desdits tableaux de référence; et

(g) identifier ledit document non identifié comme étant similaire à au moins un de ladite pluralité de documents de référence listée à l'étape quand un score à l'étape (f) est au moins égal à une valeur définie par l'utilisateur pour lesdits tableaux de référence.

L'étape de créer un tableau de référence de segments à n caractères est effectuée en :

(a) listant tous les segments distincts à n caractères qui apparaissent dans le document de référence correspondant; et

(b) affectant un poids à chaque segment à n caractères dudit document de référence.

L'étape d'affecter un poids à chaque segment à n caractères dudit document de référence est effectuée en :

(a) comptant le nombre d'occurrences de chaque segment à n caractères distinct dans ledit document de référence;

(b) divisant chaque compte de l'étape (a) par le nombre total de segments à n caractères dans ledit document de référence; et

(c) affectant chaque quotient calculé à l'étape (b) au segment à n caractères correspondant comme poids de ce segment à n caractères.

L'étape de supprimer ledit facteur commun est effectuée en :

(a) faisant une liste des segments distincts à n caractères qui apparaissent dans ladite pluralité de documents de référence ;

(b) créant un total pour chaque segment à n caractères qui représente la somme des poids de ce segment à n caractères dans ladite pluralité de documents de référence ;

(c) divisant chaque total de l'étape (b) par le nombre total de documents de référence ; et

(d) soustrayant ledit quotient de l'étape (c) du poids de son segment à n caractères correspondant dans chacun de ladite pluralité de documents de référence.

L'étape d'affecter un poids à chaque segment à n caractères dudit document non identifié est effectué en :

(a) comptant le nombre d'occurrences de chaque segment à n caractères distinct dans ledit document non identifié;

(b) divisant chaque compte de l'étape (a) par le nombre total de segments à n caractères dans ledit document non identifié ;

(c) soustrayant de chaque quotient de l'étape (b) le poids de facteur commun du segment à n caractères correspondant ; et

(d) affectant le résultat de l'étape (c) au segment à n caractères correspondant comme poids de ce segment à n caractères.

L'étape de comparer pour produire une liste de scores est effectuée en :

(a) multipliant le poids d'un segment à n caractères dans ledit document non identifié par le poids du segment à n caractères correspondant de l'un desdits tableaux de référence;

(b) répétant l'étape (a) jusqu'à ce que chaque poids de segment à n caractères dudit document non identifié ait été multiplié par le poids du segment à n caractères correspondant, s'il existe, dans le tableau référence,

(c) additionnant les produits des étapes (a) et (b) ;

(d) élevant au carré le poids de chaque segment à n caractères dans ledit document non identifié ;

(e) additionnant les carrés des poids dudit document non identifié de l'étape (d) ;

(f) élevant au carré le poids de chaque segment à n caractères dans ledit tableau de référence de l'étape (a) ;

(g) additionnant les carrés des poids dudit tableau de référence de l'étape (f) ;

(h) multipliant la somme de l'étape (e) par la somme de l'étape (g) ;

(i) prenant la racine carré du produit de l'étape (h) ;

(j) divisant la somme de l'étape (c) par le nombre calculé à l'étape (i) pour produire un score de similarité entre ledit document non identifié et ledit document de référence de l'étape (a) ; et

(k) répétant les étapes (a) à (j) pour ladite pluralité de documents de référence pour produire un score de similarité pour ledit document non identifié par rapport à ladite pluralité de documents de référence.

L'étape d'analyser est effectuée en :

(a) limitant les éléments de document possibles à un ensemble défini par l'utilisateur d'éléments de document ;

(b) remplaçant les chaînes d'éléments de document définis par l'utilisateur comprenant au moins un élément de document par des chaînes d'éléments définies par l'utilisateur comprenant au moins un élément de document ; et

(c) supprimant les chaînes d'éléments de document définies par l'utilisateur.

L'étape de limiter les éléments de document possibles à un ensemble défini par l'utilisateur d'éléments de document est effectuée en limitant les éléments de document possibles à des lettres d'une pluralité de langues.

L'étape d'identifier ledit document non identifié entraîne l'identification de la langue dudit document non identifié.

L'étape d'identifier ledit document non identifié entraîne l'identification du thème dudit document non identifié.

Pour récupérer des documents dans une langue particulière, à partir d'une base de données, par thème, l'invention comprend les étapes de :

(a) créer un tableau de référence de segments à n caractères pour chacun d'une pluralité de documents de référence qui sont dans un format lisible par machine, chaque segment à n caractères comprenant au moins un élément consécutif du document;

(b) affecter un poids à chaque segment à n caractères de l'étape (a);

(c) supprimer le facteur commun parmi ladite pluralité de tableaux de référence de l'étape (a);

(d) analyser une question qui est dans un format lisible par machine en une pluralité de segments à n caractères, chaque segment à n caractères comprenant au moins un élément de document consécutif ;

(e) affecter un poids à chaque segment à n caractères de l'étape (d);

(f) comparer chaque segment à n caractères de la question de l'étape (d) aux segments à n caractères de chacun de ladite pluralité de tableaux de référence de l'étape (a) pour produire une liste de scores qui indiquent le degré de similarité entre ladite question et chacun de ladite pluralité de documents de référence ; et

(g) identifier une pluralité de documents de référence qui sont similaires à ladite question quand un score à l'étape (f) est au moins égal à une valeur définie par l'utilisateur.

L'étape de créer un tableau de référence de segments à n caractères est effectuée en :

(a) listant tous les segments distincts à n caractères qui apparaissent dans le document de référence correspondant; et

(b) affectant un poids à chaque segment à n caractères dudit document de référence.

L'étape d'affecter un poids à chaque segment à n caractères dudit document de référence est effectuée en :

(a) comptant le nombre d'occurrences de chaque segment à n caractères distinct dans ledit document de référence ;

(b) divisant chaque compte de l'étape (a) par le nombre total de segments à n caractères dans ledit document de référence; et

(c) affectant chaque quotient calculé à l'étape (b) au segment à n caractères correspondant comme poids de ce segment à n caractères.

L'étape de supprimer ledit facteur commun est effectuée en :

(a) faisant une liste des segments distincts à n caractères qui apparaissent dans ladite pluralité de documents de référence ;

(b) créant un total pour chaque segment à n caractères qui représente la somme des poids de ce segment à n caractères dans ladite pluralité de documents de référence ;

(c) divisant chaque total de l'étape (b) par le nombre total de documents de référence ; et

(d) soustrayant ledit quotient de l'étape (c) du poids de son segment à n caractères correspondant dans chacun de ladite pluralité de documents de référence.

L'étape d'affecter un poids à chaque segment à n caractères de ladite question est effectuée en :

(a) comptant le nombre d'occurrences de chaque segment distinct à n caractères dans ladite question ;

(b) divisant chaque compte de l'étape (a) par le nombre total de segments à n caractères dans ladite question ;

(c) soustrayant de chaque quotient de l'étape (b) le poids de facteur commun du segment à n caractères correspondant ; et

(d) affectant le résultat de l'étape (c) au segment à n caractères correspondant comme poids de ce segment à n caractères.

L'étape de comparer pour produire une liste de score est effectuée en :

(a) multipliant le poids d'un segment à n caractères dans ladite question par le poids du

segment à n caractères correspondant de l'un desdits tableaux de référence;

(b) répétant l'étape (a) jusqu'à ce que tous les poids de segment à n caractères de ladite question ait été multipliés par les poids de segment à n caractères correspondants dans le tableau référence,

(c) additionnant les produits des étapes (a) et (b) ;

(d) élevant au carré le poids de chaque segment à n caractères dans ladite question ;

(e) additionnant les carrés des poids de ladite question de l'étape (d) ;

(f) élevant au carré le poids de chaque segment à n caractères dans ledit tableau de référence de l'étape (a) ;

(g) additionnant les carrés des poids dudit tableau de référence de l'étape (f) ;

(h) multipliant la somme de l'étape (e) par la somme de l'étape (g) ;

(i) prenant la racine carré du produit de l'étape (h) ;

(j) divisant la somme de l'étape (c) par le nombre calculé à l'étape (i) pour produire un score de similarité entre ladite question et ledit document de référence de l'étape (a) ; et

(k) répétant les étapes (a) à (j) pour chacun de ladite pluralité de documents de référence pour produire un score de similarité pour ladite question par rapport à chacun de ladite pluralité de documents de référence.

L'étape d'analyser est effectuée en :

(a) limitant les éléments de document possibles à un ensemble défini par l'utilisateur d'éléments de document ;

(b) remplaçant les chaînes d'éléments de document définis par l'utilisateur comprenant au moins un élément de document, par des chaînes d'éléments définis par l'utilisateur comprenant au moins un élément de document ; et

(c) supprimant les chaînes d'éléments de document définis par l'utilisateur.

L'étape de limiter les éléments de document possibles à un ensemble défini par l'utilisateur d'éléments de document est effectuée en limitant les éléments de document possibles à des lettres d'une pluralité de langues.

L'étape d'identifier une pluralité desdits documents de référence comme étant similaires à ladite question qui est dans un format lisible par machine est effectuée en identifiant une pluralité desdits documents de référence comme étant similaires à la langue de ladite question.

L'étape d'identifier une pluralité desdits documents de référence comme étant similaires à ladite question est effectuée en identifiant une pluralité desdits documents de référence comme étant similaires au thème de ladite question.

Pour classer des documents dans une base de données par langue ou par thème, l'invention comprend les étapes de :

(a) créer un tableau de référence de segments à n caractères pour chacun d'une pluralité de documents qui sont dans un format lisible par machine, chaque segment à n caractères comprenant au moins un élément consécutif d'information ;

(b) affecter un poids à chaque segment à n caractères de l'étape (a);

(c) supprimer le facteur commun parmi ladite pluralité de tableaux;

(d) comparer chaque tableau à tous les autres tableaux pour produire une liste de scores qui indiquent le degré de similarité entre chacun desdits tableaux; et

(e) classer lesdits documents en des catégories basées sur les scores à l'étape (d) et une tolérance définissable par l'utilisateur.

L'étape de créer un tableau de segments à n caractères est effectuée en :

(a) listant tous les segments distincts à n caractères qui apparaissent dans le document correspondant; et

(b) affectant un poids à chaque segment à n caractères de l'étape (a).

L'étape d'affecter un poids à chaque segment à n caractères dudit tableau est effectuée en:

(a) comptant le nombre d'occurrences de chaque segment à n caractères distinct dans ledit document;

(b) divisant chaque compte de l'étape (a) par le nombre total de segments à n caractères dans ledit document; et

(c) affectant chaque quotient calculé à l'étape (b) au segment à n caractères correspondant comme poids de ce segment à n caractères.

L'étape de supprimer ledit facteur commun est effectuée en :

(a) faisant une liste des segments distincts à n caractères qui apparaissent dans ladite pluralité de documents;

(b) créant un total pour chaque segment à n caractères qui représente la somme des poids de ce segment à n caractères dans ladite pluralité de documents;

(c) divisant chaque total par le nombre total de documents; et

(d) soustrayant ledit quotient de l'étape (c) du poids de son segment à n caractères correspondant dans chacun de ladite pluralité de documents.

L'étape de comparer pour produire une liste de scores est effectuée en :

(a) multipliant le poids d'un segment à n caractères dans l'un de ladite pluralité de documents par le poids du segment à n caractères correspondant dans un autre de ladite pluralité de documents;

(b) répétant l'étape (a) jusqu'à ce que chaque poids de segment à n caractères du premier document ait été multiplié par le poids du segment à n caractères correspondant dans le second document,

(c) additionnant les produits des étapes (a) et (b) ;

(d) élevant au carré le poids de chaque segment à n caractères dans ledit premier document;

(e) additionnant les carrés des poids dudit premier document de l'étape (d) ;

(f) élevant au carré le poids de chaque segment à n caractères dans ledit second document de l'étape (a) ;

(g) additionnant les carrés des poids dudit second document de l'étape (f) ;

(h) multipliant la somme de l'étape (e) par la somme de l'étape (g) ;

(i) prenant la racine carré du produit de l'étape (h) ;

(j) divisant la somme de l'étape (c) par le nombre calculé à l'étape (i) pour produire un score de similarité entre ledit premier document non identifié et ledit second document de l'étape (a) ; et

(k) répétant les étapes (a) à (j) pour chaque document pour produire un score de similarité pour

chaque document par rapport à chacun de ladite pluralité de documents.

Les divers objectifs de l'invention sont atteints grâce à une nouvelle approche de l'identification, de la récupération et du classement de document. Le terme "document" se rapporte à du texte, de la parole ou du graphique lisible par machine. La présente invention utilise une technique de reconnaissance de modèle basée sur des comparaisons de segments à n caractères parmi les documents, au lieu de l'approche traditionnelle par mot-clé ou fondée sur le contexte. La suppression du facteur commun parmi les documents de la base de données fournit une discrimination sensible parmi les documents, tout en permettant une réduction des besoins en mémoire (par rapport aux procédés avec mots-clés et dictionnaire) et un accroissement de performances. L'utilisateur choisit le seuil utilisé pour déterminer si des documents sont similaires.

D'autres caractéristiques et avantages de la présente invention apparaîtront plus clairement à la lecture de la description suivante de plusieurs réalisations préférées de l'invention, en référence aux dessins annexés correspondants dans lesquels :

- la figure 1 est un organigramme de l'algorithme d 'identification de langue;
- la figure 2 est un organigramme de l'algorithme d'identification de langue et de thème;
- la figure 3 est un tableau relatif à un document de référence prélevé dans la langue groenlandaise montrant tous les segments distincts à deux caractères, le nombre d'occurrence de chacun de ces segments à deux caractères et le poids de chacun des segments à deux caractères;

- la figure 4 est un tableau relatif à un document de référence prelevé dans la langue hawaïenne, montrant tous les segments distincts à deux caractères, le nombre d'occurrence de chacun de ces segments à deux caractères et le poids de chacun des segments à deux caractères;

- la figure 5 est un tableau relatif à une liste de tous les segments distincts à deux caractères issue des documents de référence des figures 3 et 4, des poids de chaque segment à deux caractères, et des poids de facteur commun que chaque segment à deux caractères montre dans les deux documents de référence;

- la figure 6A est un tableau relatif à une liste des poids diminués du facteur commun pour chaque segment à deux caractères distinct dans le document de référence groenlandais;

- la figure 6B est un tableau relatif à une liste des poids diminués du facteur commun pour chaque segment à deux caractères distinct dans le document de référence hawaïen;

- la figure 7 est un tableau relatif à un échantillon d'une langue non identifiée, montrant tous les segments distincts à deux caractères, le nombre d'occurrence de chaque segment à deux caractères, le poids de chaque segment à deux caractères, et le poids diminué du facteur commun de chaque segment à deux caractères;

- La figure 8 montre le score du texte non identifié de la figure 7 par rapport au document de référence groenlandais de la figure 3;

- La figure 9 est un organigramme de l'algorithme de récupération par langue et thème; et

- La figure 10 est un organigramme de l'algorithme de classement de la base de données par langue et thème.

La présente invention décrit un procédé pour identifier, récupérer et classer des documents. Ce procédé est un procédé sensible, rapide et économique pour l'identification de langue, l'identification de thème, la récupération de document et le classement de document. Dans toute cette description et l'objet de la présente demande de brevet défini ci-après, le terme document sera utilisé pour désigner un ensemble d'informations, dans un format lisible par machine. Le document peut consister en du texte, de la parole ou des graphiques.

Le procédé de la présente invention peut être facilement mis en oeuvre dans n'importe quel langage de programmation ou dans un matériel pour accomplir des tâches aussi difficiles que l'identification de thèmes de documents, même si ces documents sont écrits avec des idéogrammes comme le Japonais. Ce procédé peut également être utilisé pour classer des documents dans une base de données en catégories semblables sans besoin d'une identification préalable des catégories ou des documents contenus dans celles-ci.

Au lieu de compter sur des mots-clés, des tables de synonymes, ou des informations sur du contexte, les objectifs de la présente invention sont atteints par une technique de reconnaissance de modèle. La présente invention est basée sur l'hypothèse que des documents similaires en langue et/ou thème se ressemblent, en ce qu'ils ont tendance à contenir beaucoup de mêmes segments à n caractères (c'est-à-dire de chaînes consécutives à n caractères). Ils se ressemblent non pas simplement par des mots-clés mais par tous les mots employés.

Cette hypothèse suggère la conclusion plutôt étonnante que le thème d'un document peut être déduit de manière fiable en énumérant les segments à n caractères contenus dans celui-ci et en comparant le résultat de cette opération avec les segments à n caractères énumérés trouvés dans un autre document ("référence").

Cette réalisation permet des simplifications dans l'algorithme de recherche utilisé pour identifier des documents concernés. Ces simplifications ont pour résultat de permettre de classifier des documents indépendamment de la langue dans laquelle ils sont écrits. Un taux d'erreur ("altération") élevé dans le texte original est également tolérable. Le taux d'erreur tolérable maximum pour des langues dissemblables est plus élevé que le taux d'erreur tolérable maximum pour des langues similaires. Par exemple, le taux d'erreur tolérable maximum pour un texte souahéli comparé à un texte suédois peut atteindre 25% sans provoquer de résultat faux tandis que le taux d'erreur tolérable maximum pour un texte russe comparé à un texte tchèque peut seulement atteindre 15% sans provoquer de résultat faux.

La figure 1 représente un algorithme d'identification de langue. La première étape consiste à analyser le texte, qui est écrit dans une langue non identifiée, en segments à n caractères. Les segments à n caractères sont des chaînes consécutives de n caractères où n est un quelconque entier positif plus grand que zéro. Des segments à n caractères modérément longs (c'est-à-dire, n>3) sont typiquement plus porteurs d'information que des segments à n caractères plus courts, en ce qu'ils

tendent à incorporer des informations sur les racines des mots présents dans le texte. La seconde étape est de comparer le nombre d'occurrence des segments à n caractères dans le texte non identifié avec le nombre d'occurrence de ces mêmes segments à n caractères dans les textes de langues connues. Le texte non identifié est alors identifié comme étant écrit dans la langue du texte par rapport auquel le texte non identifié est le plus comparable.

Les caractères de segment à n caractères convenables sont définis par l'utilisateur. Par exemple, les éléments de segment à n caractères pour une langue particulière peuvent être restreints aux lettres de l'alphabet des langues auxquelles on s'intéresse et à un espace (c'est-à-dire "_"). Les lettres minuscules peuvent être converties en lettres majuscules et des espaces multiples peuvent être réduits à un espace simple pour réduire le nombre total de segments à n caractères possibles. La ponctuation peut également être ignorée pour minimiser les besoins en mémoire, et accroître la performance. Des chaînes de caractères peuvent aussi être éliminées ou remplacées par un caractère ou une chaîne de caractères défini par l'utilisateur.

La figure 2 représente un algorithme qui est utile pour identifier la langue ou le thème d'un document. Un exemple simple, qui explique les étapes importantes mises en jeu, suit la description de l'algorithme.

Des exemples connus de textes en différentes langues et sur différents thèmes sont recueillis en tant que comme documents de référence. Il n'y a aucune restriction sur la forme que ces documents de

référence peuvent prendre. Du texte prélevé à partir de n'importe quelle source peut être utilisé. Le nombre de documents de référence et de segments à n caractères contenus dans ces documents doit être statistiquement significatif. Il a été déterminé empiriquement que pour l'identification de langue, environ dix documents, ayant chacun environ mille caractères, est une taille d'échantillon statistiquement significative. Pour l'identification de thème dans une langue donnée, qui est une distinction plus fine que l'identification de langue, environ cinquante documents, chacun ayant environ mille caractères, est une taille d'échantillon statistiquement significative.

Les documents de référence sont analysés en segments à n caractères. Ceci est effectué en faisant une liste séparée, pour chaque document de référence, de tous les segments distincts à n caractères qui surviennent dans ce document de référence (où n est typiquement fixé à une certaine valeur qui en pratique telle que n=5). Le document non identifié est également analysé en une liste de segments distincts à n caractères.

Un poids est attribué à chacun des segments à n caractères distinct. Le poids est déterminé par le nombre relatif d'occurrence de ce segment à n caractères dans ce document de référence particulier (c'est-à-dire le nombre de fois qu'un segment à n caractères apparaît dans un document particulier de référence, divisé par le nombre total de toutes les occurrences de segments à n caractères dans ce document de référence). Des poids sont attribués à chacun des segments à n caractères dans chaque

document de référence et à chacun des segments à n caractères dans le document non identifié.

Le facteur commun parmi les documents de référence est alors supprimé des documents de référence comme du document non identifié. Ceci est effectué en listant tout d'abord les segments distincts à n caractères parmi les documents de référence. En second lieu, un poids de facteur commun est affecté à chaque segment à n caractères distinct de document de référence basé sur son nombre d'occurrence relatif moyen (c'est-à-dire la somme des poids individuels pour ce segment à n caractères dans tous les documents de référence divisée par le nombre total de documents de référence). Le poids de facteur commun de chaque segment à n caractères est alors soustrait du poids de son segment à n caractères correspondant dans chaque document de référence et du poids de son segment à n caractères correspondant dans le document non identifié.

Le document non identifié est alors comparé à chacun des documents de référence. Ceci est effectué en calculant le score du document non identifié par rapport à chacun des documents de référence. Le score du document non identifié en comparaison d'un document de référence indique le dégré de similitude entre les deux documents.

Calculer le score du document non identifié conduit tout d'abord, comme il a été mentionné ci-dessus, à soustraire le poids de facteur commun établi à partir des documents de référence, du poids de son segment à n caractères correspondant dans le document non identifié. Si un segment à n caractères apparaît dans le document non identifié sans

apparaître dans les documents de référence, le poids de facteur commun pour ce segment à n caractères est égal à zéro. Chaque poids de segment à n caractères diminué du poids de facteur commun du document non identifié (noté Ui) est alors multiplié par le poids de segment à n caractères diminué du poids de facteur commun de son segment à n caractères correspondant dans un document de référence particulier (noté Ri). Ces produits sont alors additionnés (c'est-à-dire "somme de produits").

$$\text{somme de produits} : \sum_{i=0}^{n} (Ui \times Ri)$$

Chaque poids de segment à n caractères diminué du facteur commun issu du document de référence est alors élevé au carré (c'est-à-dire Ri^2). Ces termes au carré sont alors additionnés. Chaque poids de segment à n caractères diminué du facteur commun issu de document non identifié est alors élevé au carré (c'est-à-dire Ui^2). Ces termes au carré sont également sommés. Ces deux sommes sont alors multipliées ensemble pour former un "produit de sommes" .

$$\text{produit de sommes} = \sum_{i=1}^{n} Ui^2 x \sum_{i=1}^{n} Ri^2$$

Finalement, le score du document non identifié en comparaison d'un document de référence particulier est la "somme de produits" divisée par la racine carrée du "produit de sommes" :

$$\text{score} = \sum_{i=1}^{n} (Ui \times Ri) \, x \left[\sum_{i=1}^{n} Ui^2 \, x \sum_{i=1}^{n} Ri^2 \right]^{-1/2}$$

Le document non identifié est noté par son score par rapport à chacun des documents de référence. L'utilisateur peut définir le seuil de score désiré pour identifier un document comme étant similaire à un document de référence par rapport à la langue ou au thème. Quand un seuil défini par l'utilisateur est employé, il y a une possibilité pour qu'un document non identifié ne puisse être identifié comme étant similaire à aucun des documents de référence. L'utilisateur peut éviter d'avoir à choisir ce seuil en permettant que le document non identifié soit identifié avec le document de référence qui produit le score le plus élevé.

Les exemples suivants décrivent les étapes numériques mises en jeu avec la présente invention. Ces exemples sont destinés à illustrer les étapes particulières de l'algorithme et ne doivent pas être vus comme des limitations à l'objet de la présente demande de brevet.

Le tableau dans la figure 3 montre un document de référence pour la langue groenlandaise (c'est-à-dire : "Nanok nunane issigtune). Aucun prétraitement de l'exemple n'est nécessaire. Les segments à deux caractères vont être utilisés dans cet exemple. Pour réduire le nombre de segments à deux caractères possibles, les lettres minuscules vont être rendues majuscules et chaque espace multiple va être réduit à un espace simple (c'est-à-dire "_"). Le document de référence groenlandais devient donc "NANOK_NUNANE_ISSIGTUNE".

Tous les segments distincts à deux caractères distincts pour le document de référence

groenlandais sont montrés dans le tableau de la figure 3. Ce sont NA, AN, NO, OK, K_, _N, NU, UN, NE, E_, _I, IS, SS, SI, IG, GT, TU. Un poids est affecté à chacun de ces segments à deux caractères. Le poids d'un segment à deux caractères est déterminé en divisant le nombre d'occurrence de ce segment à deux caractères particulier par le nombre total de segments à deux caractères (potentiellement non distincts) présents dans la référence (c'est-à-dire le poids du segment à deux caractères NA est 2/21 = 0,095).

Le tableau dans la figure 4 montre un document de référence pour la langue hawaïenne (c'est-à-dire : "I hele mai nei au e hai"). Tous les documents de référence doivent être analysés en utilisant la même structure de segment à n caractères (par exemple segments à deux caractères). Une fois encore, pour réduire le nombre de segments à deux caractères possibles, les lettres minuscules vont être rendues majuscules et chaque espace multiple va être réduit à un espace simple. Le document de référence hawaïen devient alors "I_HELE_MAI_NEI_AU_E_HAI".

Tous les segments distincts à deux caractères pour le document de référence hawaïen sont montrés dans la figure 4. Ce sont I_, _H, HE, EL, LE, E_, _M, MA, AI, _N, NE, EI, _A, AU, U_, _E, HA. Un poids est affecté à chaque segment à deux caractères de la même manière que décrit pour le document de référence groenlandais plus haut (par exemple, le poids du segment à deux caractères I_ est 3/22 = 0,136).

Il n'y a pas de nécessité que les nombres de segments à n caractères contenus dans les documents de référence soient semblables. L'algorithme ne demande pas de prétraitement des documents. Même la mise en majuscule et la réduction des espaces multiples à un espace simple ne sont pas nécessaires à la mise en oeuvre convenable de la présente invention. Ces étapes sont seulement effectuées dans le but d'accroître les performances et de réduire les besoins en mémoire.

Le facteur commun parmi les documents de référence est alors soustrait. Ceci est effectué en listant tout d'abord tous les segments distincts à deux caractères dans tous les documents de référence comme montré dans le tableau de la figure 5 (c'est-à-dire: NA, AN, NO, OK, K_, _N, NU, UN, NE, E_, _I, IS, SS, SI, IG, GT, TU, I_, _H, HE, EL, LE, _M, MA, AI, EI, _A, AU, U_, _E, HA). Les poids de facteur commun sont alors déterminés pour chaque segment à deux caractères distinct en divisant la somme des poids relatif à toutes les références pour chaque segment à deux caractères par le nombre total de documents de référence. Par exemple, le poids de facteur commun pour le segment à deux caractères NE est (0,095 + 0,045)/2 = 0,070 tandis que le poids de facteur commun pour le segment à deux caractères NA est (0,095 + 0)/2 = 0,048. Le segment à deux caractères NA a un poids de 0,095 dans la référence groenlandaise tandis qu'il a un poids de 0 dans la référence hawaïenne parce qu'il n'apparaît pas dans la référence hawaïenne.

Le poids de facteur commun représente la contribution moyenne d'un segment à deux caractères particulier à chacun des documents de référence. Ce

facteur commun est soustrait à la fois des documents de référence et du document non identifié pour mieux distinguer un document d'un autre. Le facteur commun est supprimé du poids d'un segment à deux caractères dans un document en soustrayant le poids de facteur commun du poids du segment à deux caractères correspondant dans ce document. Par exemple, le poids diminué du facteur commun du segment à deux caractères NE dans le document de référence groenlandais est : 0,095 - 0,070 = 0,025. Le poids diminué du facteur commun du segment à deux caractères NE dans le document de référence hawaïen est : 0,045-0,070 = - 0,025. L'étape de suppression du facteur commun améliore les performances et simplifie le processus d'identification de documents.

Le tableau dans la figure 5 liste également le poids de facteur commun de chaque segment à deux caractères distinct parmi les documents de référence. Ces poids de facteur commun sont alors soustraits des poids des segments à deux caractères correspondants dans chaque document de référence et du document non identifié. Le tableau dans la figure 6A liste les poids diminués du facteur commun pour les segments distincts à deux caractères en Groenlandais tandis que le tableau dans la figure 6B liste les poids diminués du facteur commun pour les segments distincts à deux caractères en Hawaïen. Les poids diminués du facteur commun d'un document de référence particulier sont alors utilisés pour calculer un score de similarité pour le document de référence par rapport à un document non identifié.

Le tableau dans la figure 7 montre un exemple d'un texte écrit dans une langue non identifiée (c'est-à-dire : "Martsime nanut"). Le texte non

identifié doit être analysé par la même structure de segments à n caractères que les documents de référence (c'est-à-dire, des segments à deux caractères). Encore une fois, pour réduire le nombre de segments à deux caractères possibles, les minuscules seront rendues majuscules et les espaces multiples seront réduits à un espace simple. Le document non identifié devient alors "MARTSIME_NANUT".

Tous les segments distincts à deux caractères dans le document non identifié, comme listés dans la figure 7, sont MA, AR, RT, TS, SI, IM, ME, E_, _N, NA, AN, NU, UT. Un poids est affecté à chaque segment à deux caractères. Encore une fois, le poids d'un segment à deux caractères est déterminé en divisant le nombre d'occurrence de ce segment à deux caractères par le nombre total de segments à deux caractères présents dans le texte non identifié (par exemple le poids de MA dans le texte non identifié est 1/13 = 0,077). Les poids de facteur commun des documents de référence sont alors soustraits des segments à deux caractères correspondants dans le document non identifié (voir figure 7).

Dans la figure 8, un score de similarité est calculé pour le texte non identifié par rapport au document de référence groenlandais. L'équation de ce calcul est la suivante :

$$\text{score} : \sum_{i=1}^{n} (Ui \times Ri) \times \left[\sum_{i=1}^{n} Ui^2 \times \sum_{i=1}^{n} Ri^2 \right]^{-1/2}$$

où Ui représente le poids diminué du facteur commun d'un segment à deux caractères dans le texte non identifié et Ri représente le poids diminué du facteur commun du segment à deux caractères correspondant dans le document de référence groenlandais qui est comparé au document non identifié.

Chaque poids diminué du facteur commun des segments à deux caractères du texte non identifié est multiplié par son poids diminué du facteur commun correspondant dans le document de référence groenlandais. Chaque poids de segment à n caractères diminué du facteur commun dans le document de référence groenlandais est alors élevé au carré. Ces termes au carré sont alors additionnés. Chaque poids de segment à n caractères diminué du facteur commun dans le document non identifié est alors élevé au carré. Ces termes au carré sont alors additionnés. Ces sommes sont multipliées ensemble pour former un "produit de sommes".

Finalement, le score du document non identifié par rapport au document de référence groenlandais est la "somme de produits" divisée par la racine carrée du "produit de sommes". Le résultat montré en figure 8 est un score qui indique la similarité entre le texte non identifié et le document de référence groenlandais.

Un score de similarité est généré pour chaque document de référence. Le document peut soit être identifié comme étant similaire au document de référence qui a généré le score le plus élevé ou similaire au document de référence qui a généré un score qui dépasse un seuil défini par l'utilisateur.

Dans la seconde approche, une identification n'est pas forcée (c'est-à-dire que le document non identifié peut ne pas être identifié comme similaire à l'un des documents de référence). De plus, avec la seconde approche, le document non identifié peut être identifié comme étant similaire à un grand nombre de documents de référence si ces documents de référence génèrent un score qui dépasse le seuil défini par l'utilisateur.

Certaines chaînes de texte dans une langue particulière, comme "is", "the", "and", "with", "for", etc., en Anglais, utiles pour l'identification de langue, sont normalement sans intérêt pour l'identification de thème. La présente invention résoud le problème de trouver des distinctions parmi des documents par rapport au thème en supprimant le facteur commun parmi les documents.

La suppression de facteur commun parmi les documents entraîne le calcul de la moyenne de chaque nombre d'occurrence de segment à n caractères dans tous les documents et puis la soustraction de ces moyennes de chaque nombre correspondant d'occurrence de segment à n caractères dans chaque document. La suppression de facteur commun impose l'opération de détermination si un document est similaire à un autre plus simple. La sensibilité au thème est atteinte sans intervention humaine, indépendamment de la langue du document. L'utilisateur peut encore choisir le seuil de détermination lorsque des documents sont similaires à d'autres par rapport au thème.

Cet algorithme fonctionne aussi bien quelle que soit la langue mais l'identification de thème ne fonctionne qu'en comparant des documents écrits dans

une langue commune (ou en plusieurs langues très proches). Ceci résulte de ce que le thème est lié au modèle créé par les segments à n caractères. Le modèle créé par les segments à n caractères est contrôlé par la langue du document.

L'identification de thème peut être faite aussi facilement dans des langues qui ont relativement peu de lettres, comme l'Anglais, que dans des langues qui ont beaucoup de caractères, comme les langues idéographiques (par exemple le Japonais). Bien que ne souhaitant pas relier ceci à la théorie, on croit que les résultats avantageux de la présente invention sont obtenus parce ce que le thème d'un document contraint les segments à n caractères utilisés à débattre du thème. L'identification de thème devient alors une détermination du degré de similitude des documents par rapport aux segments à n caractères utilisés.

La figure 9 indique l'algorithme qui est utilisé pour récupérer des documents d'une base de données. Les documents peuvent être récupérés par rapport à la langue ou au thème désiré.

Une base de données contient typiquement de nombreux documents concernant beaucoup de thèmes qui sont écrits en différentes langues. Il n'y a aucune nécessité que ces documents soient sous une quelconque forme particulière. Un tableau de segments à n caractères est créé pour chaque document de la base de données. Ceci est effectué en faisant une liste, pour chaque document de la base de données, des segments distincts à n caractères qui apparaissent dans ce document. Des poids sont affectés à tous les segments distincts à n

caractères. Le poids est déterminé par le nombre d'occurrence d'un segment à n caractères dans un document particulier (c'est-à-dire, le nombre de fois qu'un segment à n caractères apparaît dans un document particulier divisé par le total de toutes les occurrences de segments à n caractères dans ce document). Des poids sont alors affectés à chaque segment à n caractères dans chaque document de la base de données.

Le facteur commun parmi les documents de la base de données est alors supprimé des documents de la base de données et de la question. Ceci est accompli en listant tout d'abord tous les segments distincts à n caractères dans tous les documents. Puis un poids de facteur commun est affecté à chaque segment à n caractères distinct basé sur son nombre relatif d'occurrence moyen. Le poids de facteur commun pour chaque segment à n caractères est alors soustrait du poids de son segment à n caractères correspondant dans chaque document de la base de données et du poids de son segment à n caractères correspondant dans la question.

La question est posée par l'utilisateur. La question représente le type de document que l'utilisateur souhaite récupérer dans la base de données (c'est-à-dire des documents avec un thème ou une langue similaire). Il n'y a pas de contrainte sur la forme que doit prendre la question.

Il est à noter qu'une question concernant un thème d'intérêt aura pour résultat des documents récupérés sur ce thème et qui sont écrits dans la langue de la question. Des documents sur ce thème qui sont écrits dans une langue différente de celle de la

question ne seront normalement pas récupérés. Ceci parce que des langues différentes utilisent typiquement des segments à n caractères différents pour réprésenter le même thème.

La question est alors analysée en segment à n caractères. Ceci est effectué en faisant une liste de tous les segments distincts à n caractères qui apparaissent dans la question. Des poids sont affectés à chaque segment à n caractères distinct de la question. Le poids est déterminé par le nombre d'occurrence de chaque segment à n caractères dans la question. Les poids de facteur commun sont alors soustraits des segments à n caractères correspondants dans la question. La question est alors comparée à chacun des documents de la base de données en calculant le score de la question par rapport à chacun des documents de la base de données.

Le score est obtenu en multipliant tout d'abord chaque poids de segment à n caractères de la question diminué du facteur commun (par exemple Qi) par le poids diminué du facteur commun de son segment à n caractères correspondant dans un document particulier de la base de données (par exemple Di). Ces produits sont alors additionnés (c'est-à-dire "somme de produits").

$$\text{somme de produits} = \sum_{i=1}^{n} (Qi \times Di)$$

Chaque segment à n caractères du document de la base de données est alors élevé au carré (Di^2). Ces termes au carré sont alors additionnés. Chaque segment à n caractères de la question est alors élevé au carré (Qi^2). Ces termes au carré sont alors

additionnés. Ces sommes sont ensuite multipliées ensemble pour former "un produit de sommes".

$$\text{produit de sommes} = \sum_{i=1}^{n} Qi^2 x \sum_{i=1}^{n} Di^2$$

Finalement, le score du document de la base de données par rapport à la question est la "somme de produits" divisée par la racine carrée du "produit de sommes".

$$\text{score} = \sum_{i=1}^{n} (Qi \times Di) \times \left[\sum_{i=1}^{n} Qi^2 \times \sum_{i=1}^{n} Di^2 \right]^{-1/2}$$

Un score est généré pour chaque document de la base de données par rapport à la question. L'utilisateur définit le score-seuil qui est utilisé pour déterminer si un document de la base de données est similaire à la question par rapport à la langue ou au thème.

Les exemples donnés ci-dessus concernant les étapes mises en jeu dans l'identification de langue sont applicables pour décrire les étapes mises en jeu dans cet algorithme pour récupérer des documents depuis une base de données.

Les segments à n caractères peuvent aussi être utilisés pour résoudre les problèmes associés au classement de documents d'une base de données en catégories contenant des documents semblables. L'algorithme de la présente invention ne requiert aucune directive de la part de l'utilisateur ni aucun traitement préalable de la base de données au delà de ce qui décrit ci-dessus. La base de données peut être

classée en catégories de langue ou de thème. Des documents peuvent être facilement rattachés à des catégories différentes (ainsi, un document peut être classé dans une catégorie de langue particulière aussi bien que dans une catégorie de thème particulier).

La figure 10 décrit l'algorithme utilisé pour classer des documents d'une base de données en catégories. Les documents peuvent être classés en fonction d'une langue ou d'un thème. Une base de données contient typiquement de nombreux documents. Ces documents peuvent traiter de thèmes différents et peuvent être écrits en des langues différentes. Il n'y a aucune nécessité que ces documents soient sous une quelconque forme particulière.

Un tableau de segments à n caractères est créé pour chaque document de la base de données. Ceci est effectué en faisant une liste séparée, pour chaque document de la base de données, des segments distincts à n caractères qui apparaissent dans ce document. Des poids sont affectés à tous les segments distincts à n caractères. Le poids est déterminé par le nombre d'occurrence d'un segment à n caractères dans un document particulier. Des poids sont alors affectés à chaque segment à n caractères dans chaque document de la base de données.

Le facteur commun parmi les documents de la base de données est alors supprimé de chaque document de la base de données. Ceci est accompli en listant tout d'abord les segments distincts à n caractères qui apparaissent dans les documents (c'est-à-dire en pensant temporairement aux documents séparés de la base de données, comme à un document étendu, puis en

listant les segments distincts à n caractères qui apparaissent dans ledit document étendu). Puis un poids de facteur commun est affecté à chaque segment à n caractères distinct basé sur le nombre d'occurrence (c'est-à-dire le nombre total d'occurrence dans tous les documents de la base de données de segment à n caractères particulier divisé par le nombre total de segments distincts à n caractères dans tous les documents de la base de données). Chaque poids de facteur commun est divisé par le nombre total de documents de la base de données. Le poids de facteur commun pour chaque segment à n caractères est alors soustrait du poids de son segment à n caractères correspondant dans chaque document de la base de données.

Chaque document de la base de données est alors comparé à chacun des autres documents de la base de données. Ceci est accompli en calculant le score de chaque document de la base de données par rapport à chacun des autres documents de la base de données. Le score est obtenu en multipliant tout d'abord chaque poids de segment à n caractères diminué du facteur commun d'un document de la base de données (par exemple D1i) par le poids diminué du facteur commun de son segment à n caractères correspondant dans un document de la base de données qui lui est comparé (par exemple D2i). Ces produits sont alors additionnés pour former une "somme de produits".

$$\text{somme de produits} = \sum_{i=1}^{n} (D1i \times D2i)$$

Chaque poids de segment à n caractères diminué du facteur commun du premier document de la

base de données est alors élevé au carré ($D1i^2$). Ces termes au carré sont alors additionnés. Chaque poids de segment à n caractères diminué du facteur commun dans le second document de la base de données qui est comparé au premier document, est alors élevé au carré (par exemple $D2i^2$). Ces termes au carré sont alors additionnés. Ces sommes sont ensuite multipliées ensemble pour former "un produit de sommes".

$$\text{produit de sommes} = \sum_{i=1}^{n} D1i^2 \times \sum_{i=1}^{n} D2i^2$$

Finalement, le score du premier document de la base de données par rapport au second document de la base de données est la "somme de produits" divisée par la racine carrée du "produit de sommes".

$$\text{score} = \sum_{i=1}^{n} (D1i \times D2i) \times \left[\sum_{i=1}^{n} D1i^2 \times \sum_{i=1}^{n} D2i^2 \right]^{-1/2}$$

Un score est généré pour chaque document de la base de données par rapport à chacun des autres documents de la base de données. L'utilisateur définit le seuil que le score doit atteindre afin que deux documents qui sont comparés soient déclarés similaires. Des documents similaires sont alors classés en catégories par rapport à la langue ou au thème.

Les exemples donnés ci-dessus concernant les étapes mises en jeu dans l'identification de langue sont applicables pour décrire les étapes mises en jeu dans cet algorithme pour classer des documents d'une base de données.

Les activités de recherche actuelles en traitement de texte se sont concentrées sur une analyse basée sur le contenu (c'est-à-dire linguistique). La présente invention s'est concentrée seulement sur une analyse par reconnaissance de modèle (c'est-à-dire non linguistique). La présente invention ne requiert aucune information additionnelle concernant la synthaxe, la sémantique ou la grammaire. Les bénéfices de l'application de l'analyse par segment à n caractères à l'identification de thème n'ont pas encore été totalement appréciés jusqu'à aujourd'hui.

Le procédé d'identification de thème de la présente invention fonctionne dans n'importe quelle langue avec une facilité égale (mais dans seulement une langue à la fois). Il peut être mis en oeuvre par une personne ne lisant pas la langue du texte et ne demande pas d'apprentissage spécifique. Le procédé est robuste en présence de texte altéré (c'est-à-dire texte qui contient des erreurs). L'utilisateur choisit le seuil pour déterminer si des documents sont similaires. Le facteur commun non porteur d'information est automatiquement supprimé des documents.

Des changements et modifications dans les réalisations spécifiquement décrites, particulièrement dans le langage de programmation utilisé pour mettre en oeuvre cette invention et la formule spécifique pour le score de similarité établi à partir des poids de segment à n caractères, peuvent être apportés dans le cadre de l'objet de la présente demande de brevet.

REVENDICATIONS

1 - Procédé pour déterminer la langue d'un texte donné, caractérisé par les étapes de :

(a) analyser ledit corps de texte en une pluralité de segments à n caractères, chaque segment à n caractères comprenant le contenu de n emplacements de caractère/espace consécutifs dudit corps de texte où n a une valeur d'au moins un;

(b) comparer chacun desdits segments à n caractères à une pluralité d'ensembles de segments à n caractères où chaque ensemble de segments à n caractères représente une langue différente; et

(c) identifier la langue dudit corps de texte comme étant écrit dans la langue d'un ensemble particulier de segments à n caractères quand le pourcentage d'appariement à l'étape (b) pour un ensemble particulier de segments à n caractères est au moins égal à la valeur prescrite pour cet ensemble de segments à n caractères et plus élevé que le pourcentage d'appariement à l'étape (b) pour les autres ensembles de segments à n caractères.

2 - Procédé pour identifier un document caractérisé par les étapes de :

(a) créer un tableau de référence de segments à n caractères pour chacun d'une pluralité de documents de référence qui sont dans un format lisible par machine, chaque segment à n caractères comprenant au moins un élément consécutif du document;

(b) analyser un document non identifié qui est dans un format lisible par machine en une pluralité de segments à n caractères, chaque segment à n caractères comprenant au moins un élément consécutif de document;

(c) affecter un poids à chaque segment à n caractères de l'étape (a);

(d) affecter un poids à chaque segment à n caractères de l'étape (b);

(e) supprimer le facteur commun parmi ladite pluralité de tableaux de référence;

(f) comparer chaque segment à n caractères de document non identifié à chaque segment à n caractères desdits tableaux de référence pour produire une liste de scores qui indiquent le degré de similarité entre ledit document non identifié et chacun desdits tableaux de référence; et

(g) identifier ledit document non identifié comme étant similaire à au moins un de ladite pluralité de documents de référence listée à l'étape (a) quand un score à l'étape (f) est au moins égal à une valeur définie par l'utilisateur pour lesdits tableaux de référence.

3 - Procédé selon la revendication 2 dans lequel ladite étape de créer un tableau de référence de segments à n caractères est effectuée en :

(a) listant tous les segments distincts à n caractères qui apparaissent dans le document de référence correspondant; et

(b) affectant un poids à chaque segment à n caractères dudit document de référence.

4 - Procédé selon la revendication 3 dans lequel ladite étape d'affecter un poids à chaque segment à n caractères dudit document de référence est effectuée en :

(a) comptant le nombre d'occurrences de chaque segment à n caractères distinct dans ledit document de référence ;

(b) divisant chaque compte de l'étape (a) par le nombre total de segments à n caractères dans ledit document de référence; et

(c) affectant chaque quotient calculé à l'étape (b) au segment à n caractères correspondant comme poids de ce segment à n caractères.

5 - Procédé selon la revendication 4 dans lequel l'étape de supprimer ledit facteur commun est effectuée en :

(a) faisant une liste des segments distincts à n caractères qui apparaissent dans ladite pluralité de documents de référence ;

(b) créant un total pour chaque segment à n caractères qui représente la somme des poids de ce segment à n caractères dans ladite pluralité de documents de référence ;

(c) divisant chaque total de l'étape (b) par le nombre total de documents de référence ; et

(d) soustrayant ledit quotient de l'étape (c) du poids de son segment à n caractères correspondant dans chacun de ladite pluralité de documents de référence.

6 - Procédé selon la revendication 5 dans lequel ladite étape d'affecter un poids à chaque segment à n caractères dudit document non identifié est effectué en :

(a) comptant le nombre d'occurrences de chaque segment à n caractères distinct dans ledit document non identifié;

(b) divisant chaque compte de l'étape (a) par le nombre total de segments à n caractères dans ledit document non identifié ;

(c) soustrayant de chaque quotient de l'étape (b) le poids de facteur commun du segment à n caractères correspondant ; et

(d) affectant le résultat de l'étape (c) au segment à n caractères correspondant comme poids de ce segment à n caractères.

7 - Procédé selon la revendication 6 dans lequel ladite étape de comparer pour produire une liste de scores est effectuée en :

(a) multipliant le poids d'un segment à n caractères dans ledit document non identifié par le poids du segment à n caractères correspondant de l'un desdits tableaux de référence;

(b) répétant l'étape (a) jusqu'à ce que chaque poids de segment à n caractères dudit document non identifié ait été multiplié par le poids du segment à n caractères correspondant, s'il existe, dans le tableau référence,

(c) additionnant les produits des étapes (a) et (b) ;

(d) élevant au carré le poids de chaque segment à n caractères dans ledit document non identifié ;

(e) additionnant les carrés des poids dudit document non identifié de l'étape (d) ;

(f) élevant au carré le poids de chaque segment à n caractères dans ledit tableau de référence de l'étape (a) ;

(g) additionnant les carrés des poids dudit tableau de référence de l'étape (f) ;

(h) multipliant la somme de l'étape (e) par la somme de l'étape (g) ;

(i) prenant la racine carré du produit de l'étape (h) ;

(j) divisant la somme de l'étape (c) par le nombre calculé à l'étape (i) pour produire un score de similarité entre ledit document non identifié et ledit document de référence de l'étape (a) ; et

(k) répétant les étapes (a) à (j) pour ladite pluralité de documents de référence pour produire un score de similarité pour ledit document non identifié par rapport à ladite pluralité de documents de référence.

8 - Procédé selon la revendication 7 dans lequel ladite étape d'analyser est effectuée en :

(a) limitant les éléments de document possibles à un ensemble défini par l'utilisateur d'éléments de document ;

(b) remplaçant les chaînes d'éléments de document définis par l'utilisateur comprenant au moins un élément de document par des chaînes d'éléments définies par l'utilisateur comprenant au moins un élément de document ; et

(c) supprimant les chaînes d'éléments de document définies par l'utilisateur.

9 - Procédé selon la revendication 8 dans lequel ladite étape de limiter les éléments de document possibles à un ensemble défini par l'utilisateur d'éléments de document est effectuée en limitant les éléments de document possibles à des lettres d'une pluralité de langues.

10 - Procédé selon la revendication 9 dans lequel ladite étape d'identifier ledit document non identifié entraîne l'identification de la langue dudit document non identifié.

11 - Procédé selon la revendication 9 dans lequel ladite étape d'identifier ledit document non identifié entraîne l'identification du thème dudit document non identifié.

12 - Procédé de récupération de documents caractérisé par les étapes de :

(a) créer un tableau de référence de segments à n caractères pour chacun d'une pluralité de documents de référence qui sont dans un format lisible par machine, chaque segment à n caractères comprenant au moins un élément consécutif du document;

(b) affecter un poids à chaque segment à n caractères de l'étape (a);

(c) supprimer le facteur commun parmi ladite pluralité de tableaux de référence de l'étape (a);

(d) analyser une question qui est dans un format lisible par machine en une pluralité de segments à n caractères, chaque segment à n caractères comprenant au moins un élément de document consécutif ;

(e) affecter un poids à chaque segment à n caractères de l'étape (d) ;

(f) comparer chaque segment à n caractères de la question de l'étape (d) aux segments à n caractères de chacun de ladite pluralité de tableaux de référence de l'étape (a) pour produire une liste de scores qui indiquent le degré de similarité entre ladite question et chacun de ladite pluralité de documents de référence ; et

(g) identifier une pluralité de documents de référence qui sont similaires à ladite question quand un score à l'étape (f) est au moins égal à une valeur définie par l'utilisateur.

13 - Procédé selon la revendication 12 dans lequel ladite étape de créer un tableau de référence de segments à n caractères est effectuée en :

(a) listant tous les segments distincts à n caractères qui apparaissent dans le document de référence correspondant; et

(b) affectant un poids à chaque segment à n caractères dudit document de référence.

14 - Procédé selon la revendication 13 dans lequel ladite étape d'affecter un poids à chaque segment à n caractères dudit document de référence est effectuée en :

(a) comptant le nombre d'occurrences de chaque segment à n caractères distinct dans ledit document de référence ;

(b) divisant chaque compte de l'étape (a) par le nombre total de segments à n caractères dans ledit document de référence; et

(c) affectant chaque quotient calculé à l'étape (b) au segment à n caractères correspondant comme poids de ce segment à n caractères.

15 - Procédé selon la revendication 14 dans lequel l'étape de supprimer ledit facteur commun est effectuée en :

(a) faisant une liste des segments distincts à n caractères qui apparaissent dans ladite pluralité de documents de référence ;

(b) créant un total pour chaque segment à n caractères qui représente la somme des poids de ce segment à n caractères dans ladite pluralité de documents de référence ;

(c) divisant chaque total de l'étape (b) par le nombre total de documents de référence ; et

(d) soustrayant ledit quotient de l'étape (c) du poids de son segment à n caractères correspondant dans chacun de ladite pluralité de documents de référence.

16 - Procédé selon la revendication 15 dans lequel ladite étape d'affecter un poids à chaque segment à n caractères de ladite question est effectuée en :

(a) comptant le nombre d'occurrences de chaque segment à n caractères distinct dans ladite question ;

(b) divisant chaque compte de l'étape (a) par le nombre total de segments à n caractères dans ladite question ;

(c) soustrayant de chaque quotient de l'étape (b) le poids de facteur commun du segment à n caractères correspondant ; et

(d) affectant le résultat de l'étape (c) au segment à n caractères correspondant comme poids de ce segment à n caractères.

17 - Procédé selon la revendication 16 dans lequel ladite étape de comparer pour produire une liste de score est effectuée en :

(a) multipliant le poids d'un segment à n caractères dans ladite question par le poids du segment à n caractères correspondant de l'un desdits tableaux de référence;

(b) répétant l'étape (a) jusqu'à ce que tous les poids de segment à n caractères de ladite question ait été multipliés par les poids de segment à n caractères correspondants dans le tableau référence,

(c) additionnant les produits des étapes (a) et (b) ;

(d) élevant au carré le poids de chaque segment à n caractères dans ladite question ;

(e) additionnant les carrés des poids de ladite question de l'étape (d) ;

(f) élevant au carré le poids de chaque segment à n caractères dans ledit tableau de référence de l'étape (a) ;

(g) additionnant les carrés des poids dudit tableau de référence de l'étape (f) ;

(h) multipliant la somme de l'étape (e) par la somme de l'étape (g) ;

(i) prenant la racine carré du produit de l'étape (h) ;

(j) divisant la somme de l'étape (c) par le nombre calculé à l'étape (i) pour produire un score de similarité entre ladite question et ledit document de référence de l'étape (a) ; et

(k) répétant les étapes (a) à (j) pour chacun de ladite pluralité de documents de référence pour produire un score de similarité pour ladite question par rapport à chacun de ladite pluralité de documents de référence.

18 - Procédé selon la revendication 17 dans lequel ladite étape d'analyser est effectuée en :

(a) limitant les éléments de document possibles à un ensemble défini par l'utilisateur d'éléments de document ;

(b) remplaçant les chaînes d'éléments de document définis par l'utilisateur comprenant au moins un élément de document, par des chaînes d'éléments définis par l'utilisateur comprenant au moins un élément de document ; et

(c) supprimant les chaînes d'éléments de document définis par l'utilisateur.

19 - Procédé selon la revendication 18 dans lequel ladite étape de limiter les éléments de document possibles à un ensemble défini par l'utilisateur d'éléments de document est effectuée en limitant les éléments de document possibles à des lettres d'une pluralité de langues.

20 - Procédé selon la revendication 19 dans lequel ladite étape d'identifier une pluralité desdits documents de référence comme étant similaires à ladite question qui est dans un format lisible par machine est effectuée en identifiant une pluralité desdits documents de référence comme étant similaires à la langue de ladite question.

21 - Le procédé selon la revendication 19 dans lequel ladite étape d'identifier une pluralité desdits documents de référence comme étant similaires à ladite question est effectuée en identifiant une pluralité desdits documents de référence comme étant similaires au thème de ladite question.

22 - Procédé pour classer des documents caractérisé par les étapes de :

(a) créer un tableau de référence de segments à n caractères pour chacun d'une pluralité de documents qui sont dans un format lisible par machine, chaque segment à n caractères comprenant au moins un élément consécutif d'information ;

(b) affecter un poids à chaque segment à n caractères de l'étape (a);

(c) supprimer le facteur commun parmi ladite pluralité de tableaux;

(d) comparer chaque tableau à tous les autres tableaux pour produire une liste de scores qui

indiquent le degré de similarité entre chacun desdits tableaux; et

(e) classer lesdits documents en des catégories basées sur les scores à l'étape (d) et une tolérance définissable par l'utilisateur.

23 - Procédé selon la revendication 22 dans lequel ladite étape de créer un tableau de segments à n caractères est effectuée en :

(a) listant tous les segments distincts à n caractères qui apparaissent dans le document correspondant; et

(b) affectant un poids à chaque segment à n caractères de l'étape (a).

24 - Procédé selon la revendication 23 dans lequel ladite étape d'affecter un poids à chaque segment à n caractères dudit tableau est effectuée en:

(a) comptant le nombre d'occurrences de chaque segment à n caractères distinct dans ledit document;

(b) divisant chaque compte de l'étape (a) par le nombre total de segments à n caractères dans ledit document; et

(c) affectant chaque quotient calculé à l'étape (b) au segment à n caractères correspondant comme poids de ce segment à n caractères.

25 - Procédé selon la revendication 24 dans lequel l'étape de supprimer ledit facteur commun est effectuée en :

(a) faisant une liste des segments distincts à n caractères qui apparaissent dans ladite pluralité de documents;

(b) créant un total pour chaque segment à n caractères qui représente la somme des poids de ce segment à n caractères dans ladite pluralité de documents;

(c) divisant chaque total par le nombre total de documents; et

(d) soustrayant ledit quotient de l'étape (c) du poids de son segment à n caractères correspondant dans chacun de ladite pluralité de documents.

26 - Procédé selon la revendication 25 dans lequel ladite étape de comparer pour produire une liste de scores est effectuée en :

(a) multipliant le poids d'un segment à n caractères dans l'un de ladite pluralité de documents par le poids du segment à n caractères correspondant dans un autre de ladite pluralité de documents;

(b) répétant l'étape (a) jusqu'à ce que chaque poids de segment à n caractères du premier document ait été multiplié par le poids du segment à n caractères correspondant dans le second document,

(c) additionnant les produits des étapes (a) et (b) ;

(d) élevant au carré le poids de chaque segment à n caractères dans ledit premier document;

(e) additionnant les carrés des poids dudit premier document de l'étape (d) ;

(f) élevant au carré le poids de chaque segment à n caractères dans ledit second document de l'étape (a) ;

(g) additionnant les carrés des poids dudit second document de l'étape (f) ;

(h) multipliant la somme de l'étape (e) par la somme de l'étape (g) ;

(i) prenant la racine carré du produit de l'étape (h) ;

(j) divisant la somme de l'étape (c) par le nombre calculé à l'étape (i) pour produire un score de similarité entre ledit premier document non identifié et ledit second document de l'étape (a) ; et

(k) répétant les étapes (a) à (j) pour chaque document pour produire un score de similarité pour chaque document par rapport à chacun de ladite pluralité de documents.

FIGURE 1

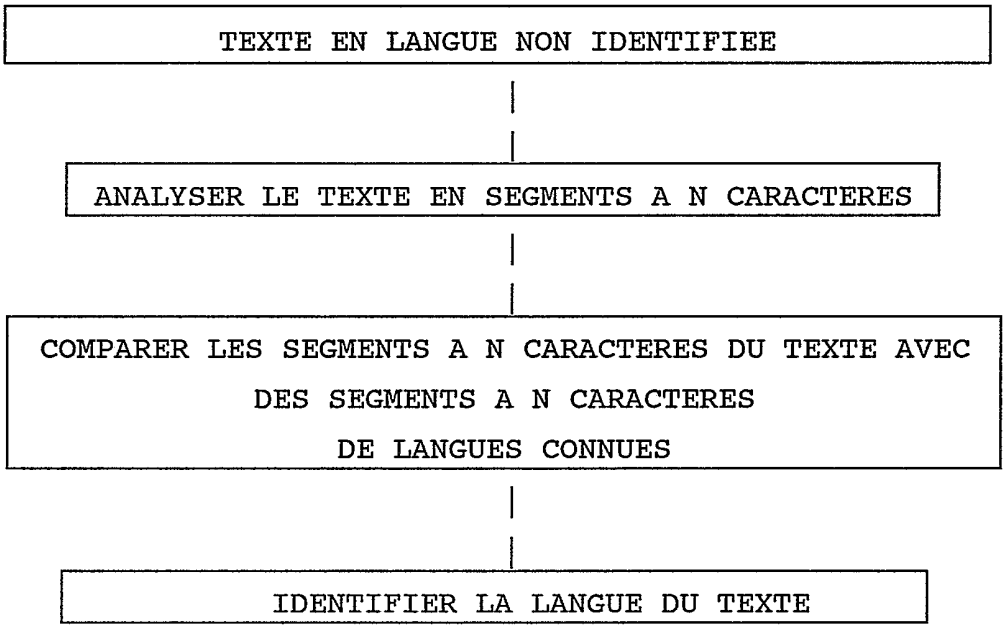

TEXTE EN LANGUE NON IDENTIFIEE
ANALYSER LE TEXTE EN SEGMENTS A N CARACTERES
COMPARER LES SEGMENTS A N CARACTERES DU TEXTE AVEC DES SEGMENTS A N CARACTERES DE LANGUES CONNUES
IDENTIFIER LA LANGUE DU TEXTE

FIGURE 2

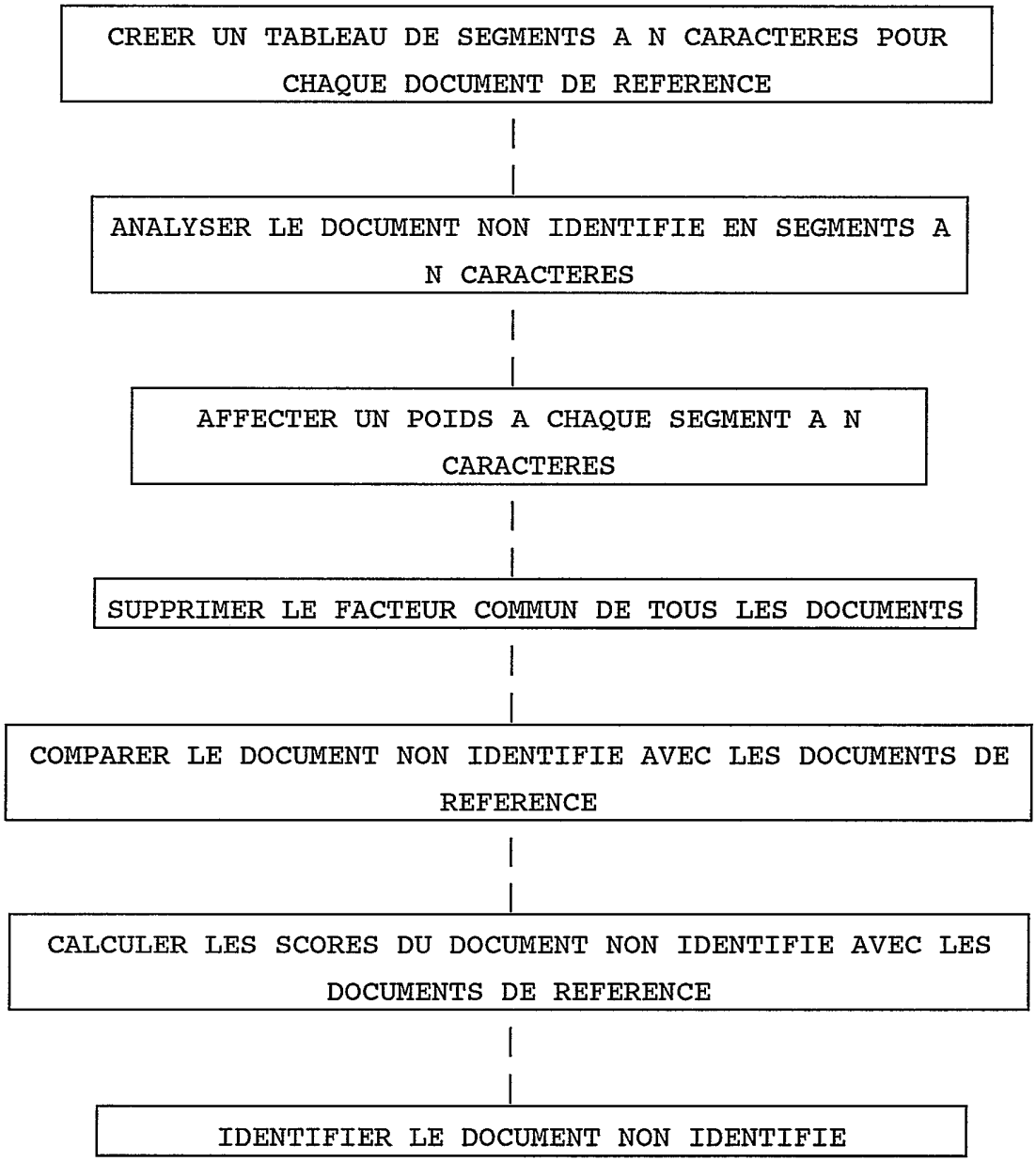

CREER UN TABLEAU DE SEGMENTS A N CARACTERES POUR CHAQUE DOCUMENT DE REFERENCE
ANALYSER LE DOCUMENT NON IDENTIFIE EN SEGMENTS A N CARACTERES
AFFECTER UN POIDS A CHAQUE SEGMENT A N CARACTERES
SUPPRIMER LE FACTEUR COMMUN DE TOUS LES DOCUMENTS
COMPARER LE DOCUMENT NON IDENTIFIE AVEC LES DOCUMENTS DE REFERENCE
CALCULER LES SCORES DU DOCUMENT NON IDENTIFIE AVEC LES DOCUMENTS DE REFERENCE
IDENTIFIER LE DOCUMENT NON IDENTIFIE

FIGURE 3

REFERENCE DE LANGUE GROENLANDAISE : NANOK_NUNANE_ISSIGTUNE

SEGMENTS DISTINCTS A_DEUX CARACTERES	NOMBRE D'OCCURRENCE	POIDS
NA	2	(2/21)=0,095)
AN	2	0,095
NO	1	0,048
OK	1	0,048
K_	1	0,048
_N	1	0,048
NU	1	0,048
UN	2	0,095
NE	2	0,095
E_	1	0,048
_I	1	0,048
IS	1	0,048
SS	1	0,048
SI	1	0,048
IG	1	0,048
GT	1	0,048
TU	1	0,048
	21	

FIGURE 4

REFERENCE DE LANGUE HAWAÏENNE : I_HELE_MAI_NEI_AU_E_HAI

SEGMENTS DISTINCTS A DEUX CARACTERES	NOMBRE D'OCCURRENCE	POIDS
I_	3	(3/22)=0,136
_H	2	0,091
HE	1	0,045
EL	1	0,045
LE	1	0,045
E_	2	0,091
_M	1	0,045
MA	1	0,045
AI	2	0,091
_N	1	0,045
NE	1	0,045
EI	1	0,045
_A	1	0,045
AU	1	0,045
U_	1	0,045
_E	1	0,045
HA	1	0,045
	22	

FIGURE 5

POIDS DE FACTEUR COMMUN POUR LES SEGMENTS A DEUX CARACTERES DANS LES DOCUMENTS DE REFERENCE

SEGMENTS DISTINCTS A DEUX CARACTERES	POIDS GROENLANDAIS	POIDS HAWAÏEN	POIDS DE FACTEUR COMMUN
NA	0,095	0,0	(0,095+0)/2=0,048)
AN	0,095	0,0	0,048
NO	0,048	0,0	0,024
OK	0,048	0,0	0,024
K_	0,048	0,0	0,024
_N	0,048	0,045	0,047
NU	0,048	0,0	0,024
UN	0,095	0,0	0,048
NE	0,095	0,045	0,070
E_	0,048	0,091	0,070
_I	0,048	0,0	0,024
IS	0,048	0,0	0,024
SS	0,048	0,0	0,024
SI	0,048	0,0	0,024
IG	0,048	0,0	0,024
GT	0,048	0,0	0,024
TU	0,048	0,0	0,024
I_	0,0	0,0136	0,068
_H	0,0	0,091	0,046
HE	0,0	0,045	0,023
EL	0,0	0,045	0,023
LE	0,0	0,045	0,023
_M	0,0	0,045	0,023
MA	0,0	0,045	0,023
AI	0,0	0,091	0,046
EI	0,0	0,045	0,023
_A	0,0	0,045	0,023
AU	0,0	0,045	0,023
U_	0,0	0,045	0,023
_E	0,0	0,045	0,023
HA	0,0	0,045	0,023

FIGURE 6A

POIDS DIMINUES DU FACTEUR COMMUN POUR LE DOCUMENT DE REFERENCE GROENLANDAIS

SEGMENTS DISTINCTS A DEUX CARACTERES	POIDS GROENLANDAIS	POIDS DE FACTEUR COMMUN	POIDS DIMINUE DU FACTEUR COMMUN
NA	0,095	0,048	(0,095-0,048)=0,047
AN	0,095	0,048	0,047
NO	0,048	0,024	0,024
OK	0,048	0,024	0,024
K_	0,048	0,024	0,024
_N	0,048	0,047	0,001
NU	0,048	0,024	0,024
UN	0,095	0,048	0,047
NE	0,095	0,070	0,025
E_	0,048	0,070	-0,022
_I	0,048	0,024	0,024
IS	0,048	0,024	0,024
SS	0,048	0,024	0,024
SI	0,048	0,024	0,024
IG	0,048	0,024	0,024
GT	0,048	0,024	0,024
TU	0,048	0,024	0,024

FIGURE 6B

POIDS DIMINUES DU FACTEUR COMMUN POUR LE DOCUMENT DE REFERENCE HAWAÏEN

SEGMENTS DISTINCTS A DEUX CARACTERES	POIDS HAWAÏEN	POIDS DE FACTEUR COMMUN	POIDS DIMINUE DU FACTEUR COMMUN
I_	0,136	0,068	(0,136-0,068)= 0,068)
_H	0,091	0,046	0,045
HE	0,045	0,023	0,022
EL	0,045	0,023	0,022
LE	0,045	0,023	0,022
E_	0,091	0,070	0,021
_M	0,045	0,023	0,022
MA	0,045	0,023	0,022
AI	0,091	0,046	0,045
_N	0,045	0,047	-0,002
NE	0,045	0,070	-0,035
EI	0,045	0,023	0,022
_A	0,045	0,023	0,022
AU	0,045	0,023	0,022
U_	0,045	0,023	0,022
_E	0,045	0,023	0,022
HA	0,045	0,023	0,022

FIGURE 7

DOCUMENT NON IDENTIFIE : MARTSIME_NANUT

SEGMENTS DISTINCTS A DEUX CARACTERES	NOMBRE D'OCCURRENCE	POIDS	FACTEURS COMMUNS	POIDS DIMINUE DU FACTEUR COMMUN
MA	1	(1/13)=0,077	0,023	0,054
AR	1	0,077	0,0	0,077
RT	1	0,077	0,0	0,077
TS	1	0,077	0,0	0,077
SI	1	0,077	0,024	0,053
IM	1	0,077	0,0	0,077
ME	1	0,077	0,0	0,077
E_	1	0,077	0,070	0,007
N	1	0,077	0,047	0,030
NA	1	0,077	0,048	0,029
AN	1	0,077	0,048	0,029
NU	1	0,077	0,024	0,053
UT	1	0,077	0,0	0,077
	13			

FIGURE 8

$$SCORE = \frac{\sum_{i=1}^{n} (Ui \times Ri)}{\sqrt{\sum_{i=1}^{n} Ui^{**2} \times \sum_{i=1}^{n} Ri^{**2}}}$$

$$SCORE = \frac{\begin{array}{l}(0.054 \times 0)+(0.077 \times 0)+(0.077 \times 0)+(0.077 \times 0)+ \\ (0.053 \times 0.024)+(0.077 \times 0)+(0.077 \times 0)+ \\ (0.007 \times (-0.022))+ (0.030 \times 0.001)+(0.029 \times 0.047)+ \\ (0.029 \times 0.047)+(0.053 \times 0.024)+(0.077 \times 0)\end{array}}{\sqrt{\begin{array}{l}((0.054^{**}2)+(0.077^{**}2)+(0.077^{**}2)+(0.077^{**}2)+ \\ (0.053^{**}2)+(0.077^{**}2)+(0.077^{**}2)+(0.007^{**}2)+ \\ (0.030^{**}2)+(0.029^{**}2)+(0.029^{**}2)+(0.053^{**}2)+ \\ (0.077^{**}2)) \times ((0.047^{**}2)+(0.047^{**}2)+(0.024^{**}2)+ \\ (0.024^{**}2)+(0.024^{**}2)+(0.001^{**}2)+(0.024^{**}2)+ \\ (0.047^{**}2)+(0.025^{**}2)+(-0.022^{**}2)+(0.024^{**}2)+ \\ (0.024^{**}2)+(0.024^{**}2)+(0.024^{**}2)+(0.024^{**}2)+ \\ (0.024^{**}2)+(0.024^{**}2))\end{array}}}$$

FIGURE 9

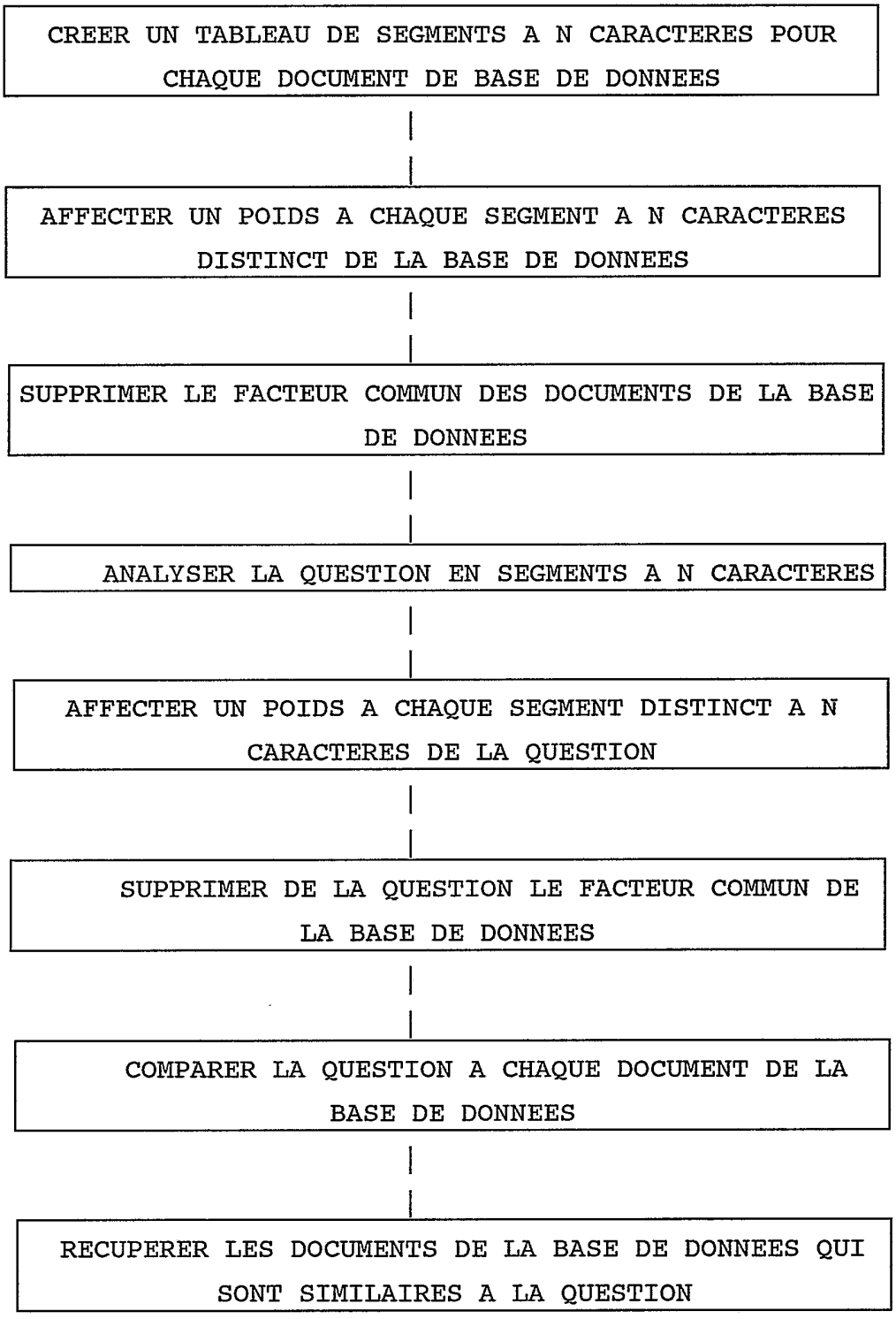
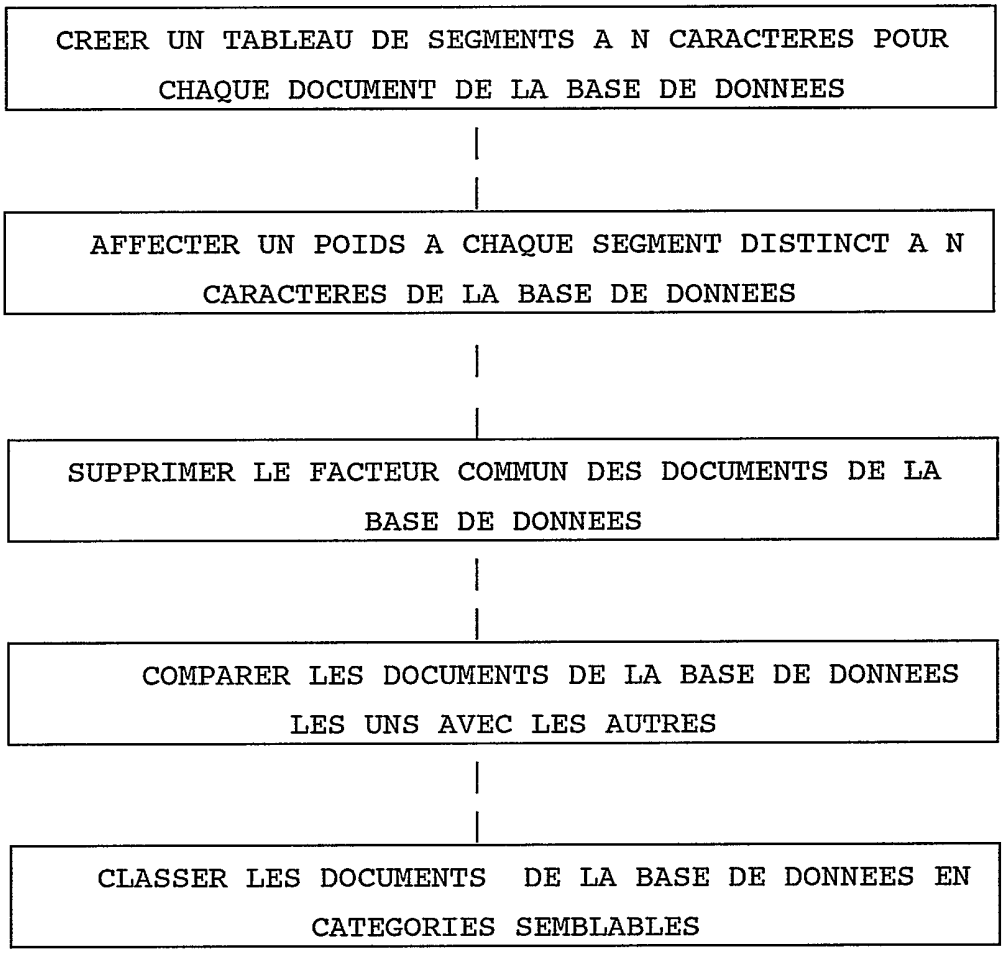

CREER UN TABLEAU DE SEGMENTS A N CARACTERES POUR CHAQUE DOCUMENT DE BASE DE DONNEES
AFFECTER UN POIDS A CHAQUE SEGMENT A N CARACTERES DISTINCT DE LA BASE DE DONNEES
SUPPRIMER LE FACTEUR COMMUN DES DOCUMENTS DE LA BASE DE DONNEES
ANALYSER LA QUESTION EN SEGMENTS A N CARACTERES
AFFECTER UN POIDS A CHAQUE SEGMENT DISTINCT A N CARACTERES DE LA QUESTION
SUPPRIMER DE LA QUESTION LE FACTEUR COMMUN DE LA BASE DE DONNEES
COMPARER LA QUESTION A CHAQUE DOCUMENT DE LA BASE DE DONNEES
RECUPERER LES DOCUMENTS DE LA BASE DE DONNEES QUI SONT SIMILAIRES A LA QUESTION

FIGURE 10


CREER UN TABLEAU DE SEGMENTS A N CARACTERES POUR CHAQUE DOCUMENT DE LA BASE DE DONNEES
AFFECTER UN POIDS A CHAQUE SEGMENT DISTINCT A N CARACTERES DE LA BASE DE DONNEES
SUPPRIMER LE FACTEUR COMMUN DES DOCUMENTS DE LA BASE DE DONNEES
COMPARER LES DOCUMENTS DE LA BASE DE DONNEES LES UNS AVEC LES AUTRES
CLASSER LES DOCUMENTS DE LA BASE DE DONNEES EN CATEGORIES SEMBLABLES

REPUBLIQUE FRANÇAISE

INSTITUT NATIONAL
de la
PROPRIETE INDUSTRIELLE

RAPPORT DE RECHERCHE

établi sur la base des dernières revendications déposées avant le commencement de la recherche

2694984

N° d'enregistrement national

FR 9211714
FA 476623

DOCUMENTS CONSIDERES COMME PERTINENTS		Revendications concernées de la demande examinée
Catégorie	Citation du document avec indication, en cas de besoin, des parties pertinentes	
X	US-A-5 062 143 (J. C. SCHMITT) * abrégé; revendications; figures 2,3 *	1-10, 12-20-22 -26
X	2ND INTERNATIONAL CONFERENCE ON COMPUTERS AND APPLICATIONS 23 Juin 1987, PEKIN, CHINE pages 65 - 69 A. MELTZER & G KOWALSKI : 'Text searching using an inversion database consisting of trigrams' * abstract * * le document en entier *	1-10, 12-20, 22-26
X	IBM TECHNICAL DISCLOSURE BULLETIN. vol. 24, no. 11A, Avril 1982, NEW YORK US pages 5402 - 5403 F. DAMERAU : 'Variable n-gram method for statistical language processing' * le document en entier *	1-10, 12-20, 22-26

DOMAINES TECHNIQUES RECHERCHES (Int. Cl.5)

G06F

Date d'achèvement de la recherche
07 JUIN 1993

Examinateur
FOURNIER C.D.J.

CATEGORIE DES DOCUMENTS CITES

X : particulièrement pertinent à lui seul
Y : particulièrement pertinent en combinaison avec un autre document de la même catégorie
A : pertinent à l'encontre d'au moins une revendication ou arrière-plan technologique général
O : divulgation non-écrite
P : document intercalaire

T : théorie ou principe à la base de l'invention
E : document de brevet bénéficiant d'une date antérieure à la date de dépôt et qui n'a été publié qu'à cette date de dépôt ou qu'à une date postérieure.
D : cité dans la demande
L : cité pour d'autres raisons

& : membre de la même famille, document correspondant

EPO FORM 1503 03.82 (P0413)