(19) CONFÉDÉRATION SUISSE
INSTITUT FÉDÉRAL DE LA PROPRIÉTÉ INTELLECTUELLE

(11) CH 717 260 A2

(51) Int. Cl.: G06F 16/38 (2019.01)
G06F 16/338 (2019.01)
G06N 3/08 (2006.01)
G06F 16/332 (2019.01)

Demande de brevet pour la Suisse et le Liechtenstein
Traité sur les brevets, du 22 décembre 1978, entre la Suisse et le Liechtenstein

(12) DEMANDE DE BREVET

(21) Numéro de la demande: 00364/20

(22) Date de dépôt: 26.03.2020

(43) Demande publiée: 30.09.2021

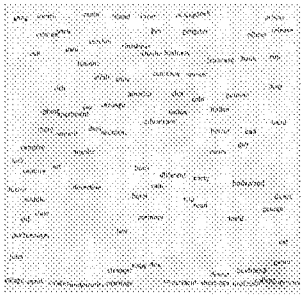
(71) Requérant:
NeoInstinct SA Artificial Common Sense,
Rue Traversière 3
1018 Lausanne (CH)

(72) Inventeur(s):
François Blayo, 1814 La Tour-de-Peilz (CH)

(74) Mandataire:
BOVARD SA Conseils en propriété intellectuelle,
Optingenstrasse 16
3013 Bern (CH)

(54) Méthode mise en oeuvre par ordinateur pour la recherche analogique de documents.

(57) La présente invention se rapport à une méthode mise en oeuvre par ordinateur pour la recherche analogique de documents comprenant au moins une information textuelle d'un ensemble de documents E compris dans une première base de données et dont l'information textuelle correspond le plus à un terme de recherche R comprenant les étapes :
a. Génération d'une deuxième base de données comprenant une liste de mots produite par lemmatization de l'information textuelle des documents de la première base de données ;
b. Génération d'un vecteur descripteur V de valeurs numériques pour chaque document de la première base de données à l'aide d'une fonction de vectorisation F de l'information textuelle ;
c. Apprentissage d'une carte auto-organisée C comprenant un réseau de P neurones p par projection des vecteurs descripteurs V sur la carte auto-organisée C, chaque neurone p de la carte auto-organisée C correspondant à un vecteur poids W de valeurs numériques ;
d. Allocation de chaque document de la première base de données au neurone p de la carte auto-organisée C dont le vecteur poids W correspondant possède la plus petite distance avec le vecteur descripteur V du document à allouer;
e. Génération à l'aide de la fonction de vectorisation F et de la deuxième base de données d'un vecteur de recherche K de valeurs numériques pour le terme de recherche R ;
f. Détermination du neurone pbest de la carte auto-organisée C dont le vecteur poids W possède la plus petite distance avec le vecteur de recherche K; et
g. Détermination des documents de la première base de données alloués au neurone pbest de la carte auto-organisée C.



Description

Domaine technique de l'invention

[0001] La présente invention se rapporte au domaine des méthodes mises en oeuvre par ordinateur pour la recherche de documents. Plus précisément, la présente invention concerne une méthode mise en oeuvre par ordinateur pour la recherche analogique de documents dans un ensemble de documents. Elle concerne spécifiquement une méthode mise en oeuvre par ordinateur qui permet d'identifier des documents dans un ensemble de documents qui sont rapprochent le plus d'un terme de recherche qui n'est pas forcément contenu dans les documents de l'ensemble de documents.

État de la technique

[0002] L'évolution rapide de la numérisation documentaire a conduit à une production explosive de fichiers contenant une information codée de manières très variées : documents textes, traitements de texte, documents sous format PDF et en général tout document dans un format capable de contenir des informations textuelles. L'exploitation efficace de ces documents passe par la possibilité de trouver rapidement un contenu recherché par un utilisateur. De nombreuses solutions ont été apportées à ce besoin et notamment des moteurs de recherche mais aussi des systèmes de gestion documentaire ou des systèmes d'indexation.

[0003] Les méthodes de recherche conventionnelles utilisent une série de mots ou de termes clés appariés à un corpus d'information connu, par exemple le contenu des documents indexés. Un utilisateur saisit un mot ou une expression de recherche et un moteur de recherche fait correspondre le mot ou l'expression à une liste de mots dérivés du corpus. Le moteur de recherche affiche ensuite les documents qui correspondent au mot ou à l'expression en question. Cependant, les méthodes de recherche conventionnelles ne tiennent pas compte des deux principaux problèmes inhérents à la recherche.

[0004] Le premier problème se rapporte à une recherche imprécise. Dans une recherche imprécise, un utilisateur entre des informations incorrectes dans la requête. Si l'utilisateur n'entre pas les bons mots dans la requête, les documents extraits ne représenteront pas l'intention de l'utilisateur. Par exemple, si un utilisateur entre un mot de recherche particulier dans une requête, mais que l'information désirée est indexée sous un synonyme du mot de recherche, alors l'utilisateur ne retrouvera pas l'information désirée.

[0005] Le deuxième problème se rapporte à une recherche vague. Dans une recherche vague, l'utilisateur n'en sait pas assez sur le sujet de la l'information désirée pour former une requête de recherche précise. Par exemple, si l'utilisateur ne connaît pas le langage pour décrire une condition médicale particulière, alors l'utilisateur ne sera pas en mesure d'entrer correctement la requête de recherche.

[0006] Dans ce contexte, la possibilité d'une recherche analogique peut apporter un avantage à un utilisateur qui souhaite trouver une information dans une grande masse de documents tout en formulant une requête imprécise ou vague.

[0007] Le principe de la recherche par analogie a été proposé notamment dans le contexte de l'image où des solutions proposent de rechercher par similitude. Néanmoins, dans le contexte de la recherche textuelle, la recherche analogique soulève la question de déterminer ce qu'est une analogie entre plusieurs documents.

[0008] Quelques solutions ont été proposées pour la recherche analogique dans le contexte de la recherche textuelle. En particulier, le brevet US 888 657 9 B2 concerne le traitement sémantique du texte par les réseaux de neurones, c'est-à-dire l'analyse du sens d'un texte en se concentrant sur la relation entre ses mots et ce qu'ils représentent dans le monde réel et dans leur contexte.

[0009] Le brevet US9183288B2 présente une méthode efficace de structurer les données pour une recherche efficace et fiable en exploitant les relations sémantiques entre les documents. Il utilise une techniques d'analyse sémantique pour créer un modèle d'espace vectoriel de documents contenus dans un corpus de domaine puis crée une structure hiérarchique des documents par le biais d'un processus d'agglomération. Chaque document dans un corpus de domaine est mis en correspondance avec les autres documents dans le même corpus de domaine pour déterminer quels documents sont les plus similaires.

[0010] Bien que les solutions connues de l'art antérieur permettent d'identifier des documents dans un ensemble de documents à l'aide d'une recherche analogique, ces solutions ne sont pas entièrement satisfaisantes. En effet, ces solutions permettent une mise en relation des documents mais ne permettent pas une exploration visuelle basée sur la proximité sémantique des documents. Notamment, ces solutions ne présentent pas les documents sous forme de carte de proximité contenant les termes les plus représentatifs des régions de cette carte.

[0011] Il existe par conséquent un besoin pour une méthode mise en oeuvre par ordinateur qui permet à l'aide d'une recherche analogique d'identifier des documents d'un ensemble de documents qui correspondent le plus à un terme de recherche. Le but de de la présente invention consiste en particulier à offrir une méthode mise en oeuvre par ordinateur permettant de résoudre les problèmes tels que :

- accéder le plus simplement et naturellement à une base de données documentaire constituée de nombreux documents de contenu très variés ;

– suggérer à l'utilisateur des documents qui ne correspondent pas explicitement à ses requêtes mais dont le contenu peut éventuellement l'inspirer ;
– représenter le contenu de la base de données documentaire sous forme d'une carte résumant son contenu ; et
– améliorer le processus et améliorer les réponses aux demandes en tenant compte des interactions possibles avec un utilisateur.

Résumé de l'invention

[0012] Un but de la présente invention est donc de proposer une méthode mise en oeuvre par ordinateur pour la recherche analogique de documents textuels permettant de surmonter les limitations mentionnées préalablement.

[0013] Selon l'invention, ces buts sont atteints grâce aux objets de la revendication indépendante. Les aspects plus spécifiques de la présente invention sont décrits dans les revendications dépendantes ainsi que dans la description.

[0014] De manière plus spécifique, un but de l'invention est atteint grâce à une méthode mise en oeuvre par ordinateur pour la recherche analogique de documents textuels d'un ensemble de documents E compris dans une première base de données et dont le contenu correspond le plus à un terme de recherche R comprenant les étapes :

a. Génération d'une deuxième base de données comprenant une liste de mots produite par lemmatization des documents de la première base de données ;

b. Génération d'un vecteur descripteur V de valeurs numériques pour chaque document de la première base de données à l'aide d'une fonction de vectorisation F de l'information textuelle ;

c. Apprentissage d'une carte auto-organisée C comprenant un réseau de P neurones p sur la base des vecteurs descripteurs V, chaque neurone p de la carte auto-organisée C correspondant à un vecteur poids W de valeurs numériques ;

d. Allocation de chaque document de la première base de données au neurone p de la carte auto-organisée C dont le vecteur poids W correspondant possède la plus petite distance avec le vecteur descripteur V du document à allouer;

e. Génération à l'aide de la fonction de vectorisation F et de la deuxième base de données d'un vecteur de recherche K de valeurs numériques pour le terme de recherche R ;

f. Détermination du neurone pbest de la carte auto-organisée C dont le vecteur poids W possède la plus petite distance avec le vecteur de recherche K ; et

g. Détermination des documents de la première base de données alloués au neurone pbest de la carte auto-organisée C.

[0015] Grâce à une telle méthode, il est possible d'identifier le ou les documents d'un ensemble de documents dont le contenu se rapproche le plus d'un terme de recherche. Il est particulier possible d'accéder simplement et naturellement à une base de données documentaire constituée de nombreux documents de contenu très variés, de suggérer à l'utilisateur des documents qui ne correspondent pas explicitement à ses requêtes mais dont le contenu peut éventuellement l'inspirer. En d'autres termes, le résultat de la recherche se fonde sur la ressemblance du terme de recherche des documents de l'ensemble de documents. Les documents les plus „ressemblants“ peuvent être identifiés même si le terme de recherche n'est pas contenu dans aucun des documents de l'ensemble de documents. A noter que l'apprentissage de la carte auto-organisée C à l'étape c. peut être mis en oeuvre par tout algorithme d'apprentissage non-supervisé connus de l'art antérieur, tel que par exemple, l'algorithme dit de Kohonen.

[0016] Dans un tel algorithme, la carte auto-organisée C est composée d'un réseau de faible dimension de P neurones p. Quand le réseau est unidimensionnel, chaque neurone possède deux voisins. Quand le réseau est bidimensionnel, l'arrangement des neurones se fait d'une façon rectangulaire où chaque neurone possède quatre voisins (topologie rectangulaire) ou d'une façon hexagonale où chaque neurone possède six voisins (topologie hexagonale). Les neurones sont reconnus par leur numéro et leur emplacement dans le réseau.

[0017] Les vecteurs descripteurs V sont projetés de leur espace initial, ou espace d'entrée, vers la carte ou espace de sortie. A chaque neurone p de la carte est associé un vecteur poids W, appelé aussi vecteur prototype ou référent, appartenant à l'espace d'entrée. En désignant par P le nombre total des neurones de la carte, le vecteur poids W du neurone p de dimension N est désigné par :
W_p avec $p \in \{1, ..., P\}$ et $W_p \in R^N$

[0018] L'objectif de l'apprentissage de la carte consiste à mettre à jour les vecteurs poids de façon à approximer au mieux la distribution des vecteurs descripteurs V tout en reproduisant l'auto-organisation des neurones de la carte. L'apprentissage de la carte se fait en mode séquentiel appelé aussi incrémental, ou en mode différé (batch).

[0019] Chaque itération t de l'apprentissage séquentiel comprend deux étapes. La première étape consiste à choisir au hasard un vecteur descripteur V(t) de l'ensemble Ω, et à le présenter au réseau dans le but de déterminer son neurone vainqueur. Le neurone vainqueur (Best Matching Unit), d'une observation est le neurone dont le vecteur poids en est le plus proche au sens d'une distance donnée (par exemple une distance euclidienne). Si c est le neurone vainqueur du vecteur V(t), c est déterminé comme suit :

$$d\big(W_c(t), V(t)\big) = \min_{p\in\{1,\dots,P\}} d\left(W_p(t), V(t)\right)$$

[0020] Dans une deuxième étape, le neurone vainqueur est activé. Son vecteur poids West mis à jour pour se rapprocher du vecteur descripteur présenté au réseau. Cette mise à jour ne concerne pas seulement le neurone vainqueur comme dans les méthodes de l'apprentissage par compétition (Winner take ail), mais aussi les neurones qui lui sont voisins et qui voient alors leurs vecteurs poids s'ajuster vers ce vecteur descripteur. L'amplitude de cet ajustement est avantageusement déterminée par la valeur d'un pas d'apprentissage α(t) et la valeur d'une fonction de voisinage h(t).

[0021] Le paramètre α(t) règle la vitesse de l'apprentissage. Il est initialisé avec une grande valeur au début puis décroît avec les itérations en vue de ralentir au fur et à mesure le processus d'apprentissage. La fonction h(t) définit l'appartenance au voisinage. Elle dépend à la fois de l'emplacement des neurones sur la carte et d'un certain rayon de voisinage. Dans les premières itérations, le rayon de voisinage est assez large pour mettre à jour un grand nombre de neurones voisins du neurone vainqueur, mais ce rayon se rétrécit progressivement pour ne contenir que le neurone vainqueur avec ses voisins immédiats, ou bien même le neurone vainqueur seulement. La règle de mise à jour des vecteurs poids est la suivante :
$W_p(t+1) = W_p(t) + \alpha(t)h_{cp}(t)[V(t)-W_p(t)]$ avec $p \in (1,\dots,P)$
où c est le neurone vainqueur du vecteur descripteur V(t) présenté au réseau à l'itération t et h est la fonction de voisinage qui définit la proximité entre les neurones c et p.

[0022] Une fonction de voisinage entre le neurone vainqueur c et un neurone p de la carte vaut 1 si le neurone p se trouve à l'intérieur du carré centré sur le neurone c et 0 dans les autres cas. Le rayon de ce carré est appelé rayon de voisinage. Il est large au début, puis se rétrécit avec les itérations pour contenir seulement le neurone c avec ses voisins immédiats à la fin de l'apprentissage ou même seulement le neurone c. Une fonction de voisinage plus flexible et plus commune est la fonction gaussienne définie ci-dessous :

$$h_{cp}\big(\sigma(t)\big) = exp\left(-\frac{\|r_c - r_p\|^2}{2\sigma^2(t)}\right)$$

où r_c et r_p sont respectivement l'emplacement du neurone c et du neurone p sur la carte, et σ(t) est le rayon du voisinage à l'itération t du processus d'apprentissage.

[0023] Avec une telle fonction de voisinage, l'amplitude de l'ajustement est graduée selon l'éloignement du neurone vainqueur qui réserve à lui-même l'amplitude maximale. Le résultat de l'apprentissage non supervisé de la carte auto-organisée C est la projection non linéaire de l'ensemble des vecteurs descripteurs V sur la carte. Chaque vecteur descripteur V est attribué à son neurone vainqueur ce qui permet d'allouer chaque documents de l'ensemble de documents à un neurone de la carte auto-organisée. Outre la tâche de quantification, cette projection préserve la topologie des données grâce à l'utilisation de la fonction de voisinage. Deux neurones voisins sur la carte représenteront des observations proches dans l'espace de données.

[0024] Une variante de l'apprentissage est dite „en mode différé". En mode différé, à chaque itération t, tous les vecteurs descripteurs V sont présentés au réseau et la mise à jour des vecteurs poids se fait en prenant en compte tous les vecteurs descripteurs V. Chaque vecteur poids est une moyenne pondérée des vecteurs descripteurs (V_i, $i \in \{1, \dots, n\}$). Quand le carré de la distance euclidienne est utilisé pour le calcul du neurone vainqueur, les poids correspondants sont les valeurs de la fonction de voisinage h(t).

[0025] La règle de mise à jour des vecteurs prototypes est donnée par :

$$W_p(t+1) = \frac{\sum_{i=1}^{n} h_{pc_i}(t)V_i}{\sum_{i=1}^{n} h_{pc_i}(t)} \quad \text{avec } p \in \{1, \dots, P\}$$

où h est la valeur de la fonction de voisinage entre le neurone vainqueur c_i du vecteur V_i et le neurone p.

[0026] La mise à jour des vecteurs prototypes peut être formulée autrement en utilisant le fait que les observations qui ont le même neurone vainqueur ont la même valeur pour la fonction de voisinage et appartiennent à la région de Voronoï dont le centre est leur neurone vainqueur :

$$W_p(t+1) = \frac{\sum_{l=1}^{K} h_{pl}(t)n_l\overline{V_l}}{\sum_{l=1}^{K} n_l h_{pl}(t)} \quad \text{avec } p \in \{1, \dots, P\}$$

où n_l est le nombre de vecteurs descripteurs V appartenant à la région de Voronoï représentée par le neurone l et V_l est la moyenne des vecteurs descripteurs de cette même région.

[0027] Vers la fin de l'apprentissage, quand le rayon de voisinage devient trop petit pour activer seulement le neurone vainqueur, chaque vecteur poids constitue le centre de gravité des observations qu'il représente et on retombe alors sur l'algorithme des centres-mobiles, ce qui garantit une meilleure approximation de la fonction de densité des observations. De plus, avec l'absence du pas d'apprentissage, cet algorithme ne présente pas de problèmes de convergence.

[0028] Cependant, le mode différé pourrait causer des torsions dans les cartes à grandes dimensions. Pour cette raison, on procède à une analyse en composantes principales pour initialiser les vecteurs prototypes.

[0029] De manière avantageuse, la carte auto-organisée est une carte bidimensionnelle ou tridimensionnelle. L'initialisation de la carte C avant la procédure d'apprentissage en tant que telle peut être effectuée de plusieurs façons. Par exemple, une première méthode d'initialisation consiste à assigner un vecteur de poids W initial à chaque noeud de la carte auto-organisée C. Cette attribution initiale des vecteurs de poids peut être par exemple une attribution aléatoire d'un nombre à chaque vecteur scalaire des vecteurs poids, sans stimulation. Le terme „aléatoire" désigne probabilité égale pour n'importe lequel d'un ensemble de résultats possibles. La valeur numérique de ces valeurs scalaires assignées au hasard peut être approximativement limitée à la borne inférieure et supérieure par l'extrema correspondant observé dans les vecteurs descripteurs V. Une autre méthode d'initialisation des vecteurs poids W comprend une variation systématique, par exemple une variation linéaire, dans la plage de chaque dimension de chaque vecteur de poids pour recouper approximativement la plage correspondante observée dans les vecteurs descripteurs V. Dans une autre méthode d'initialisation, les vecteurs poids W sont initialisés par les valeurs des vecteurs ordonnés le long d'un sous-espace bidimensionnel traversé par les deux vecteurs propres principaux des vecteurs descripteurs V obtenus par des méthodes d'orthogonalisation bien connues dans l'art, par exemple par l'orthogonalisation dite de Gram-Schmidt. Dans une autre procédure d'initialisation, les valeurs initiales sont fixées sur des échantillons choisis au hasard parmi les vecteurs descripteurs V.

[0030] La détermination du neurone vainqueur de la carte auto-organisée C pour chaque vecteur descripteur V peut se faire selon plusieurs critères bien connus de l'homme du métier. Cela peut par exemple être effectué sur la base, d'une distance par exemple la distance Euclidienne minimale entre tous les vecteurs poids W de la carte auto-organisée C et le vecteur V. D'autres méthodes peuvent pour la détermination du neurone vainqueur telles que celles utilisant la corrélation entre vecteurs qui présente l'avantage d'offrir plus de robustesse au décalage entre vecteurs, l'écart angulaire entre vecteurs qui offre l'avantage de mettre l'accent sur la longueur mutuelle des vecteurs pour autant que l'information soit portée par ces grandeurs, la mesure de distance de Minkowsky qui est une généralisation de la mesure de distance euclidienne et qui est avantageuse lorsque les vecteurs portent des données de nature qualitatives peuvent être aussi mises en oeuvre.

[0031] Dans un premier mode de réalisation préféré de la présente invention, la distance entre deux vecteurs est une distance euclidienne. La distance Euclidienne entre vecteurs est une mesure qui peut être déterminée très rapidement quelle que soit la dimension de la carte auto-organisée C ce qui permet une mise en oeuvre rapide de la présente méthode et donc une recherche rapide du ou des documents dont le contenu ressemble le plus au terme de recherche. De plus, la détermination d'une distance Euclidienne entre deux vecteurs ne demande que peu de ressources de calcul. Elle peut donc se faire sur des ordinateurs de bureau ordinaires.

[0032] Dans un autre mode de réalisation préféré de la présente invention, le contenu textuel des documents de la première base de données est normalisé avant la génération des vecteurs descripteurs V. Le procédé de normalisation est couramment utilisé par toute personne connaissant l'état de l'art en matière de prétraitement de documents textuels. Les opérations typiquement réalisées lors de la normalisation sont, de manière non exhaustive, l'agrégation de mots, la transformation des lettres majuscules en lettres minuscules pour les noms communs, la suppression de caractères de ponctuation, la transformation des pronoms de liaison. La normalisation permet de supprimer l'information redondante ou inutile dans le texte des documents

[0033] Dans un autre mode de réalisation préféré de la présente invention, la fonction de vectorisation F est une fonction dite de „term frequency-inverse document frequency". La méthode dite „TF-IDF „(de l'anglais term frequency-inverse document frequency) est une méthode de pondération souvent utilisée en recherche d'information et en particulier dans la fouille de textes. Cette mesure statistique a l'avantage de permettre une évaluation de l'importance d'un terme contenu dans un document, relativement à une collection ou un ensemble de documents. L'importance, aussi appelé poids, augmente proportionnellement au nombre d'occurrences du mot dans le document et varie également en fonction de la fréquence du mot dans l'ensemble de documents.

[0034] Dans un autre mode de réalisation préféré de la présente invention, la carte auto-organisée C est bi- ou tridimensionnelle. Une carte bi- ou tridimensionnelle permet de réduire la complexité des calculs à effectuer lors de la recherche sans perte trop importante d'information. Elles permettent en outre de produire un résultat de recherche qui regroupe plusieurs documents de manière simple tout en conservant leur proximité sémantique.

[0035] Dans encore un autre mode de réalisation préféré de la présente invention, la carte auto-organisée C est affichable graphiquement par l'intermédiaire d'une interface graphique qui est par exemple l'interface graphique d'un ordinateur personnel. Grâce à ceci le contenu de la carte C est accessible par l'intermédiaire d'une interface graphique et peut être

exploré directement par un utilisateur. En particulier, le contenu de la première base de données peut être affiché sur la carte auto-organisée. Ceci permet de voir la proximité des différents documents contenus dans la première base de données. Ainsi si, par exemple, un document a été identifié comme étant particulièrement pertinent, il est possible de trouver les documents dont le contenu est semblable à celui-ci très facilement car ils sont positionnés proches sur la carte auto-organisée C.

[0036] Dans encore un autre mode de réalisation préféré de la présente invention, une représentation graphique d'une carte de documents CD de dimension égale à la carte auto-organisée C est superposable à la représentation graphique de la carte auto-organisée C, et les documents déterminés à l'étape g. sont identifiés graphiquement sur la carte de documents CD. Ceci permet de créer un deuxième layer, ou deuxième couche, dans l'interface graphique. Ce deuxième layer qui est superposable à la représentation graphique de la carte auto-organisée comprend une représentation graphique des documents identifiés comme étant les plus proches du terme de recherche R.

[0037] Dans un mode de réalisation préféré de la présente invention suivant, le contenu d'un document déterminé à l'étape g. est accessible en sélectionnant ce document sur la représentation graphique de la carte de documents CD par l'intermédiaire d'un dispositif de pointage pour ordinateur, tel que par exemple une souris informatique. Ceci permet d'accéder très rapidement au contenu des documents les plus proches du terme de recherche R.

[0038] Dans encore un autre mode de réalisation préféré de la présente invention, une représentation graphique d'une carte de distances CH de dimension égale à la carte auto-organisée C est superposable à la représentation graphique de la carte auto-organisée C, et une valeur de distance dd est attribuée à chaque neurone de la carte de distances CH, la valeur de distance dd correspondant à la somme des distances euclidienne entre le vecteur poids W du neurone considéré et les vecteurs poids W des neurones voisins directs. Comme mentionné auparavant, l'algorithme d'apprentissage de la carte auto-organisée C a pour effet de regrouper les documents proches au sens d'une mesure de distance dans des neurones. Les neurones sont codés dans une matrice sous forme de vecteurs de données réelles. Ces neurones sont ordonnés par l'algorithme de telle sorte que des documents proches dans l'espace des données soient aussi proches que possibles sur la carte C. Néanmoins, à cause de la forte non-linéarité de la projection des données d'origine, c'est-à-dire les vecteurs descripteurs V, sur la carte C, la proximité entre les neurones ne donne aucune information sur la distance réelle qui sépare les vecteurs descripteurs V dans l'espace d'origine. Il en résulte que des documents alloués à des neurones proches sur la carte C peuvent en réalité correspondre à des données très distantes dans l'espace d'origine et donc en réalité être très différents. Cette limitation peut être en partie réduite par l'utilisation de la carte de distances CH qui peut être superposée à la carte auto-organisée C dans l'interface graphique. Cette carte de distances CH est une carte de dimension égale à la carte C et qui donne une mesure de la distance réelle entre les vecteurs poids W de cette dernière. Cette mesure peut être avantageusement affichée sur une représentation graphique de la carte de distances CH par une „lookup table“ adaptée, par exemple une coloration adaptée ou un niveau de gris adapté.

[0039] Dans encore un autre mode de réalisation préféré de la présente invention, une couleur d'un codage couleur est attribuée à chaque neurone de la carte de distance CH en fonction de la valeur dd. Ceci permet de déterminer visuellement très rapidement le niveau de ressemble réel de deux documents positionnés de façon proches sur la carte auto-organisée C.

[0040] Dans un mode de réalisation préféré de la présente invention suivant, une représentation graphique d'une carte de mots CW de dimension égale à la carte auto-organisée C est superposable à la représentation graphique de la carte auto-organisée C, dans laquelle à chaque neurone de la représentation graphique de la carte de mots CW sont affichés les mots du vocabulaire dont la composante du vecteur poids W du neurone de la carte auto-organisé C correspondant est supérieure à une valeur prédéterminée. Tout comme une carte routière représente les villes dans leur contexte incluant les monuments, les routes, les forêts et en général tout ce qui donne un contexte à la ville, la représentation des documents sur une carte peut être contextualisée en plaçant sur la carte de mots CW les mots les plus significatifs à proximité des neurones de la carte C correspondant aux documents qui contiennent ces mots. Une représentation graphique de la carte de mots CW est avantageusement superposée à la représentation graphique de la carte auto-organisée C afin de donner une information supplémentaire à l'utilisateur.

[0041] Dans encore un autre mode de réalisation préféré de la présente invention, les mots affichés dans la carte de mots CW sont organisés sur des plans différents correspondant à des différentes plages de valeurs de poids, les mots dont la valeur de poids sont les plus élevées étant affichées au premier plan de la représentation graphique de la carte de mots CW. Le nombre de mots à afficher sur la représentation graphique de la carte CW peut être très élevé et conduire à un affichage illisible si ces derniers étaient tous placés sur un même plan. Dans le cadre de cette invention, l'affichage des documents sur la carte est avantageusement allégé en offrant un système de zoom comparable à celui qui est disponible pour les cartes routières. Pour ce faire, les mots sont distribués sur différents plans d'affichage selon leur pertinence.

[0042] Dans encore un autre mode de réalisation préféré de la présente invention, les documents de la première base de données sont indexés dans une troisième base données, et les documents comprenant le terme de recherche R sont déterminés. Ceci permet en plus de la recherche analogique de faire une recherche sur la base d'une indexation des documents de la première base de données. Il est ainsi possible de rechercher et d'identifier les documents qui contiennent explicitement le terme de recherche R.

[0043] Dans encore un mode de réalisation préféré de la présente invention suivant, les documents comprenant le terme de recherche R sont identifiés sur la représentation graphique de la carte de documents CD. Ceci permet d'avoir rapidement une information quant à la ressemblance de deux documents qui ont été identifiés par la recherche textuelle. En effet, deux documents proches sur la carte de documents CD ont un contenu proche. Bien sûr la carte de distance CH peut dans ce cas aussi être superposée à la carte de documents CD. Cela permet de fournir à l'utilisateur une indication sur la ressemblance réelle de deux documents identifiés lors de la recherche textuelle.

[0044] Dans encore un autre mode de réalisation préféré de la présente invention, les documents de la première base de données sont des documents codés sous forme numérique avantageusement issus de traitement de texte, de systèmes de reconnaissance de textes, par exemple par l'intermédiaire de la méthode dite Optical Character Recognition, ou issus de tout système capable de produire des fichiers numériques structurés. Ceci a l'avantage que le contenu textuel des documents peut être facilement traité pour créer le vocabulaire ainsi que les vecteurs descripteurs.

Brève description des dessins

[0045] Les particularités et les avantages de la présente invention apparaîtront avec plus de détails dans le cadre de la description qui suit avec un exemple de réalisation donné à titre illustratif et non limitatif en référence aux dix-neuf figures ci-annexées qui représentent :

- La figure 1 représente un schéma fonctionnel d'une méthode selon un mode de réalisation de la présente invention ;
- La figure 2 illustre l'étape d'apprentissage de la carte auto-organisée C ;
- La figure 3 représente un schéma fonctionnel de l'adaptation de la carte auto-organisée ;
- La figure 4 illustre le système de génération des cartes documents, mots et distances;
- La figure 5 montre une représentation graphique de la carte de documents ;
- La figure 6 représente un schéma fonctionnel de la recherche des documents les plus proches du terme de recherche ;
- La figure 7 montre une représentation graphique de la carte de distances ;
- La figure 8 illustre la relation entre la carte, neurones, les poids de mots et le vocabulaire ;
- La figure 9a illustre la première étape dans le calcul des coordonnées continues pour un mot du vocabulaire;
- La figure 9b illustre la deuxième étape dans le calcul des coordonnées continues pour un mot du vocabulaire;
- La figure 10 illustre le positionnement d'un mot spécifique sur la carte de mots après calcul des coordonnées continues ;
- La figure 11 illustre la distribution des mots sur différents plans de la carte de mots ;
- La figure 12 montre exemple de positionnement de mots sur une carte de 10'000 neurones ;
- La figure 13 montre l'interface graphique ;
- La figure 14 montre la superposition des cartes de mots, de distances et de documents sur la carte auto-organisée ;
- La figure 15 illustre l'accès à un résumé d'un document par sa sélection sur la carte de documents ;
- La figure 16 montre l'interface graphique avec le résumé des documents sélectionnés sur la carte de documents ;
- La figure 17 illustre le système d'indexation et de recherche textuelle de documents ;
- La figure 18 montre un exemple d'un résultat d'une recherche analogique dans l'interface graphique ;

Description détaillée d'un mode de réalisation

[0046] L'invention présentée ici consiste à permettre d'identifier un ou plusieurs documents parmi un ensemble de documents à l'aide d'une recherche analogique.

[0047] En résumé, l'invention emploie un système d'analyse de documents déposés dans un emplacement de stockage centralisé. L'analyse des documents produit automatiquement un vocabulaire utilisé pour coder les documents sous forme de vecteurs descripteurs caractéristiques. Ces vecteurs descripteurs sont cartographiés sous forme d'une carte auto-organisée, de préférence bidimensionnelle, qui peut être enfin utilisée pour effectuer les recherches dans la base des documents et identifier un ou plusieurs documents correspondant le plus à un terme de recherche.

[0048] L'invention repose sur la possibilité de cartographier les documents sur une carte auto-organisée, par exemple une carte auto-organisée bidimensionnelle, de telle manière que deux documents dont les vecteurs descripteurs sont proches dans l'espace des vecteurs descripteurs sont placés sur la carte à des emplacements proches. Cette propriété présente l'avantage d'effectuer des regroupements de documents à partir de leur seul contenu et sans supervision. Finalement, un utilisateur pourra avantageusement exploiter par l'intermédiaire d'une interface graphique appropriée la carte auto-organisée afin de rechercher et trouver des documents par simple exploration de la carte. La cartographie repose ainsi sur une projection non-linéaire de points depuis l'espace constitué par les vecteurs descripteurs de documents vers une carte en deux dimensions.

[0049] Un schéma général d'un mode de réalisation préféré de l'invention est présenté dans la Figure 1. Dans une première étape, un système de lecture de documents 110 prend en entrée des documents textuels, d'un ensemble de documents E, codés sous forme numérique qui peuvent être issus de traitement de texte, de systèmes de reconnaissance de textes, par exemple par l'intermédiaire de la méthode connue sous l'abréviation OCR (Optical Character Récognition) et en général tout système capable de produire des fichiers numériques structurés.

[0050] Le contenu textuel des documents de l'ensemble de documents E est enregistré dans une première base de données, appelées ici base de données „documents bruts“ 120 qui est utilisée comme source du système de traitement de document 130 dont l'objet est, dans une deuxième étape, de traiter les contenus extraits des documents pour regrouper des mots semblables et supprimer les caractères de ponctuation.

[0051] Les documents ainsi traités sont stockés dans la base de données „documents normalisés“ 140 qui est utilisée, dans une troisième étape, par le système de génération de vocabulaire 150 pour produire une liste de mots appelée „vocabulaire“ établie en appliquant des restrictions à la liste de mots produite par le système de traitement de documents 130. Le vocabulaire ainsi obtenu est stocké dans une deuxième base de données, ici appelée base de données „vocabulaire“ 160, qui va être utilisé, dans une quatrième étape, par le système de génération de vecteurs descripteurs de documents 170 qui transforme chaque document de la base de données 140 en vecteurs descripteurs de documents qui sont finalement stockés dans la base de données „vecteurs descripteurs“ 175.

[0052] Les vecteurs descripteurs stockés dans la base de données 175 sont utilisés, dans une cinquième étape, par le système de génération et de traitement de cartes de documents 180 qui produit une carte auto-organisée C qui permet la recherche analogique d'un ou plusieurs documents de l'ensemble de documents E sur la base d'un terme de recherche R avantageusement définit par un utilisateur par l'intermédiaire d'une interface graphique 190 qui est avantageusement une interface graphique d'un ordinateur personnel, tel que par exemple un écran d'ordinateur. Cette interface graphique 190 permet l'affichage graphique des documents identifiés lors de la recherche sur la carte auto-organisée C et sur une ou plusieurs cartes supplémentaires (voir ci-dessous pour plus de détails).

[0053] Dans ce mode de réalisation non limitatif, parallèlement au de flux de traitement des documents décrit ci-dessus et permettant la recherche analogique, un système d'indexation de documents 125 traite les documents enregistrés dans la base de données 120 et indexe leur contenu qui est enregistré dans une troisième base de données appelées ici base de données „d'indexation“. Cette indexation peut être avantageusement utilisée pour permettre, en plus du mode de recherche analogique, un mode de recherche textuel qui sera disponible dans l'interface graphique 190.

[0054] L'interface graphique 190 est avantageusement utilisée pour visualiser la carte auto-organisée C et une ou plusieurs cartes supplémentaires, les agrandir, les déplacer, afficher ou cacher des mots, afficher ou cacher des documents, et pour rechercher les documents par deux modes de recherche analogique et textuel. Les différentes étapes de la méthode schématisée dans la Figure 1 vont maintenant être décrites en détails grâce à un exemple concret d'application de la présente invention.

[0055] Le système de lecture de document 110 représente tout dispositif capable de lire des documents textuels et de les enregistrer dans une base de données 120. Chaque ligne de la base de données 120 contient pour chaque document un numéro ID et le contenu du document sous forme de texte brut. Le Tableau 1 monte un exemple d'un contenu typique d'une ligne de la base de données „documents bruts“ 120 qui est obtenue à partir de la base de données de 44'512 résumés de „l'Internet Movie Database“.

ID	Contenu du document
19	A vengeful New York transit cop décidés to steal a trainload of subway fares; his foster brother, a fellow cop, tries to protect him.

Tableau 1 : une ligne de la base de données „documents bruts“ 120

[0056] Le système de traitement des documents 130 prend en entrée les informations contenues dans la base de données „documents bruts“ 120 pour effectuer une série d'analyses et de transformations destinées à normaliser le contenu des documents sous une forme qui permettra son exploitation ultérieurement. Ce procédé appelé „normalisation“ est couramment utilisé par toute personne connaissant l'état de l'art en matière de prétraitement de documents textuels. Les opérations typiquement réalisées lors de la normalisation sont, de manière non exhaustive, l'agrégation de mots, la transformation des lettres majuscules en lettres minuscules pour les noms communs, la suppression de caractères de ponctuation, la transformation des pronoms de liaison et en général tout traitement visant à supprimer l'information redondante ou inutile dans le texte des documents. Pour la langue anglaise, la normalisation effectuera, par exemple, les transformations suivantes :

- happier, happiest seront transformés en happy
- worse, worst seront transformés en badly
- dogs, children seront transformés en dog, child
- writes, writing, wrote, written seront transformés en write
- les pronoms seront transformés en -PRON-

[0057] La ligne de la base de données „documents bruts“ 120 présentée dans le Tableau 1 sera ainsi transformée et codée dans la base de données „documents normalisés“ 140 par la ligne présentée dans le Tableau 2.

ID	Contenu du document
19	a vengeful New York transit cop décidé to steal a trainload of subway fare - PRON- foster brother a fellow cop try to protect -PRON-

Tableau 2. une ligne de la base de données „documents normalisés“ 140

[0058] Après exécution du système de traitement de documents 130, les documents sont ainsi normalisés et peuvent être, dans une étape suivante, utilisés pour construire le „vocabulaire“.

[0059] Le vocabulaire est un ensemble de mots sélectionnés parmi tous les mots contenus dans les documents de la base de données „documents normalisés“ 140. Il a pour objet de représenter de manière concise toute l'information textuelle des documents sous une forme canonique. En ce sens, le vocabulaire constitue les axes d'un espace multidimensionnel dont la dimension est égale au nombre de mots du vocabulaire.

[0060] L'étape de génération du vocabulaire est connue par les personnes qui maîtrisent l'état de l'art sous le nom de „lemmatization“ car elle consiste à générer des „lemmes“ ou mots de vocabulaire. Dans le mode de réalisation présenté ici, la génération du vocabulaire va consister entre autres à compter tous les mots qui apparaissent dans tous les documents de la base de données „documents normalisés“ 140.

[0061] Une manière habituelle pour construire le vocabulaire consiste à partir d'un vocabulaire existant et à compter le nombre de mots qui apparaissent dans les documents normalisés. Une autre méthode consiste à construire dynamiquement le vocabulaire à partir des documents normalisés. En effet, les mots sont par construction éligibles à devenir des mots de vocabulaire grâce au prétraitement effectué lors du traitement de documents 130. Dans le mode de réalisation présenté ici, deux décomptes seront effectués (1) le nombre d'apparitions de chaque mot dans l'ensemble des documents normalisés ainsi que (2) le nombre de documents normalisés dans lesquels ce mot apparaît. De plus, deux paramètres seront choisis arbitrairement lors de la construction du vocabulaire :

- VOCABULARY _CHOICE_DOCCOUNT_MIN_DOCS : le nombre minimum de documents dans lesquels les mots apparaissent ; et
- VOCABULARY _CHOICE_DOCCOUNT_MAX_DOCS: le nombre maximum de documents dans lesquels les mots apparaissent.

[0062] Ces deux paramètres ont pour objectif de réduire la taille du vocabulaire en supprimant les mots qui apparaissent trop souvent ainsi que ceux qui apparaissent trop rarement dans l'ensemble des documents. A l'issue de l'exécution du système de génération de vocabulaire 150, la base de données „vocabulaire“ 160 contiendra autant de lignes que de mots du vocabulaire, chacun d'eux étant représenté par 3 valeurs :

- Le mot (word)
- Le nombre d'apparition de ce mot dans tous les documents (count)
- Le nombre de documents dans lesquels ce mot apparaît (documents)

[0063] A titre d'exemple, six lignes de la base de données „vocabulaire“ 160 correspondant à l'exemple concret utilisé ici sont présentées dans le Tableau 3.

word	count	documents
join	1025	999
gang	1273	999
many	1068	993
write	1037	979
arrive	1004	976
part	1115	971

Tableau 3: extrait de six lignes de la base de données „vocabulaire“ 160

[0064] Il est important de noter que toute méthode capable de produire un vocabulaire à partir des documents normalisés de la base de données „documents normalisés“ 140 pourrait être utilisée dans le cadre de la présente invention et son choix dépendra des avantages et inconvénients que la méthode présente.

[0065] Les vecteurs descripteurs de documents V sont des vecteurs dont les composantes sont des données numériques qui sont calculées à partir des données textuelles de documents stockées dans la base de données „documents normalisés“ 140 ainsi que de la base de données „vocabulaire“ 160 qui a été construite à l'issue de l'exécution des systèmes 110, 130 et 150.

[0066] De très nombreuses méthodes existent pour transformer des textes en valeurs numériques. Citons par exemple le comptage de mots dans un document, qui délivre une seule valeur numérique. Cette méthode présente l'avantage d'être très simple à réaliser mais résulte en une perte importante d'information sur le contenu du document. Une autre méthode pourrait consister à définir arbitrairement des mots-clefs et à compter le nombre d'apparitions de ces mots-clefs dans chaque document. Cette méthode présente l'avantage de capter une information plus riche dans chaque document mais présente l'inconvénient de devoir construire une liste de mots-clefs arbitraire.

[0067] Une autre méthode possible et connue sous le nom de „bag of words“ consiste à compter la fréquence d'apparition d'un vocabulaire de mots dans chaque document et de construire un vecteur de nombres qui contient la fréquence calculée pour chaque mot du dictionnaire. Cette méthode apporte une richesse encore plus élevée mais requiert des calculs importants et surtout ne donne aucune information sur la fréquence d'apparition relative d'un mot dans l'ensemble des documents qui a servi à construire le vocabulaire.

[0068] La méthode dite „TF-IDF „(de l'anglais term frequency-inverse document frequency) est une méthode de pondération souvent utilisée en recherche d'information et en particulier dans la fouille de textes. Cette mesure statistique permet d'évaluer l'importance d'un terme contenu dans un document, relativement à une collection ou un ensemble de documents. Le poids augmente proportionnellement au nombre d'occurrences du mot dans le document et varie également en fonction de la fréquence du mot dans l'ensemble de documents. C'est cette méthode qu'il est utilisé dans le mode réalisation préféré de l'invention présenté ici. Cette méthode peut être toutefois être remplacée par toute autre méthode standard ou originale qui pourrait apporter une information plus adaptée au domaine d'application relatif aux documents traités sans sortir du cadre de la présente invention.

[0069] Dans la méthode TF-IDF, la fréquence „brute“ TF d'un terme correspond simplement au nombre d'occurrences de ce terme dans le document considéré. A noter que le terme de „fréquence“ est un abus de langage. Le terme de „fréquence“ sera toutefois utilisé ici, car il est régulièrement utilisé dans le domaine technique de la présente invention. Il est possible de choisir cette fréquence brute pour exprimer la fréquence d'un terme. Dans ce cas, le calcul de la fréquence brute s'exprime par :
$TF_i = f_{i,d}$ où f représente la fréquence brute, i est le mot considéré et d est le document considéré.

[0070] La fréquence inverse de document IDF est une mesure de l'importance du terme dans l'ensemble des documents. Dans le schéma TF-IDF, elle vise à donner un poids plus important aux termes les moins fréquents, considérés comme plus discriminants.

[0071] En général, la détermination de la fréquence inverse IDF consiste à calculer l'inverse de la proportion de documents de l'ensemble qui contiennent le terme :

$$IDF_i = log \frac{|D|}{|\{d_j : t_i \in d_j\}|}$$

Où :
$|D|$: le nombre total de documents dans l'ensemble de documents ;
$|\{d_j: t_i \in d_j\}|$: le nombre de documents où le mot t_i apparaît.

[0072] La valeur TF-IDF pour un couple TF - $IDF_{i,j}$ est donné par :
$TF\text{ - }IDF_{i,j} = TF_{i,j} \times IDF_i$

[0073] Toujours en utilisant notre exemple „a vengeful New York transit cop décidé to steal a trainload of subway fare -PRON- foster brother a fellow cop try to protect -PRON-“ de la base de données „documents normalisés“ 140, la valeur TF-IDF pour le mot „cop“ dans le document avec l'ID 19 du Tableau 2 sera calculée de la manière suivante :
- Pour l'exemple, on suppose que tous les mots de la phrase font partie du vocabulaire et que le mot „cop“ apparaît dans 5 documents sur 100 documents au total.

$$TF_{cop,19} = \frac{2}{24}$$

$$IDF_{cop} = \log\frac{100}{5}$$

et donc:

$$TF_{cop,19}\ x\ IDF_{cop} = \frac{2}{24} x\ \log\frac{100}{5} = 0.108$$

[0074] Ainsi, à la fin de l'exécution du système de génération de descripteurs de documents 170, chaque document sera codé par un vecteur descripteur V dont le nombre de composantes correspond au nombre de mots dans le vocabulaire. Les composantes du vecteur descripteur V de chaque document résultent quant à elles du calcul de TF-IDF décrit ci-dessus. Chaque ligne de la base de données „vecteurs descripteurs" 175 contiendra les valeurs TF-IDF associées à chaque document stocké dans la base de données „documents bruts" 120 et suivant le vocabulaire stocké dans la base de données „vocabulaire" 160.

[0075] Une dernière opération consiste, avantageusement, en une normalisation de la matrice des vecteurs descripteurs V contenue dans la base de données „vecteurs descripteurs" 175 en appliquant une normalisation dite „L2", appelée aussi norme euclidienne. Lors d'une normalisation L2 les valeurs sont normalisées de sorte que si elles étaient toutes mises au carré et additionnées, le total serait égal à 1.

[0076] En se référant à nouveau à l'exemple concret, si le vocabulaire stocké dans la base de données „vocabulaire" 160 contient 4'096 mots, chaque document de la base de données „documents normalisés" 140 sera codé dans la base de données „vecteurs descripteurs" 175 comme un vecteur descripteur V de valeurs réelles de dimension 4'096, chaque valeur résultant du calcul du TF-IDF de chaque mot du vocabulaire pour chaque document. Chaque ligne de la base de données „vecteurs descripteurs" 175 représente ainsi un vecteur descripteur V de chaque document. Ces vecteurs descripteurs V seront utilisés ultérieurement pour générer une carte auto-organisée C des documents avec le système de génération et de traitement de la carte de documents 180.

[0077] Par exemple, la ligne du Tableau 2 pour le document avec l'ID 19 sera codée par le vecteur descripteur V montré dans le

Colonne 13	Colonne 99	Colonne 223	Colonne 323	Colonne 1399	Colonne 2319	Colonne 3065
0.2626657 7532894786	0.5520229 829217238	0.2906555 087191916	0.3017655 135684061	0.3617619 7580917313	0.390883 868973761	0.4087448 356906496
steal	cop	protect	fellow	foster	vengeful	subway

[0078] Tableau 4. Seules les sept colonnes non vides sont représentées pour un vocabulaire contenant 4'096 mots, obtenu à partir du traitement des 44'512 résumés de films extraits de „l'Internet Movie Database".

Colonne 13	Colonne 99	Colonne 223	Colonne 323	Colonne 1399	Colonne 2319	Colonne 3065
0.2626657 7532894786	0.5520229 829217238	0.2906555 087191916	0.3017655 135684061	0.3617619 7580917313	0.390883 888973761	0.4087448 356906496
steal	cop	protect	fellow	foster	vengeful	subway

Tableau 4: valeurs du vecteur descripteur V pour le document 19 du Tableau 2

[0079] Le système de génération et de traitement de carte de documents 180 est destiné à produire une carte auto-organisée C qui regroupe tous les documents contenus dans la base de données „documents normalisés" 140 sur forme d'une carte qui place les documents dont le contenu est similaire à des emplacements proches sur cette carte. Pour ce faire, les données stockées dans la base de données „vecteurs descripteurs" 175 sont utilisées pour alimenter un système de classification automatique.

[0080] Le système de génération de carte 180 utilise avantageusement l'algorithme dit des „Self-Organizing Maps (SOM)" qui produit une carte auto-organisée C comme cela est illustré sur la Figure 2.

[0081] La carte auto-organisée C est composée d'une grille de neurones p de faible dimension. Quand la grille est unidimensionnelle, chaque neurone p a deux voisins. Quand la grille est bidimensionnelle, l'arrangement des neurones p se fait d'une façon rectangulaire où chaque neurone possède quatre voisins (topologie rectangulaire) ou d'une façon hexagonale où chaque neurone possède six voisins (topologie hexagonale). Les neurones p sont identifiés par leur numéro et leur emplacement sur la grille.

[0082] Les vecteurs descripteurs de documents V= v(1), v(2), ... ,v(p) sont projetés de leur espace initial, ou espace d'entrée, vers la carte auto-organisée C ou espace de sortie. A chaque neurone p de la carte C est associé un vecteur poids W, appelé aussi vecteur poids ou prototype, appartenant à l'espace d'entrée. En désignant par P le nombre total des neurones p de la carte C, le vecteur poids W du neurone p de dimension N est désigné par :
W_p avec $p \in \{1, ..., P\}$ et $W_p \in R^N$

[0083] L'objectif de l'apprentissage de la carte consiste à mettre à jour les vecteurs poids W de façon à approximer au mieux la distribution des vecteurs d'entrée, c'est-à-dire les vecteurs descripteurs V, tout en reproduisant l'auto-organisation des neurones p de la carte C. L'apprentissage de la carte peut se faire avantageusement en mode séquentiel appelé aussi incrémental, ou en mode différé (batch). Le processus général de l'apprentissage est décrit sur la Figure 3.

[0084] Tous les vecteurs poids W sont initialisés à des valeurs aléatoires à l'étape 810. Chaque itération t de l'apprentissage séquentiel comprend deux étapes. La première étape consiste à choisir au hasard un vecteur descripteur V(t) de l'ensemble des vecteurs descripteurs contenu dans la base de données „vecteurs descripteurs" 175 (étape 820), et à le présenter au réseau de neurones p dans le but de déterminer son neurone vainqueur (étape 830). Le neurone vainqueur, appelé Best Matching Unit ou BMU, d'un vecteur descripteur V(t) est le neurone p dont le vecteur poids W(t) en est le plus proche au sens d'une distance donnée, par exemple la distance euclidienne. Si c est le neurone vainqueur, c'est-à-dire le BMU du vecteur descripteur V(t), c est déterminé comme suit :

$$d\big(W_c(t), V(t)\big) = \min_{p\in\{1,\ldots,P\}} d\Big(W_p(t), V(t)\Big)$$

[0085] Dans la deuxième étape d'une itération t, le neurone vainqueur est activé. Son vecteur poids W(t) est mis à jour pour se rapprocher du vecteur descripteur V(t) présenté au réseau. Cette mise à jour ne concerne pas seulement le neurone vainqueur BMU, comme dans les méthodes d'apprentissage par compétition dites de „winner take all", mais aussi les neurones qui lui sont voisins et qui voient alors leurs vecteurs poids W(t) s'ajuster également vers le vecteur descripteur V(t). L'amplitude de cet ajustement 840 est déterminée par la valeur d'un pas d'apprentissage $\alpha(t)$ et la valeur d'une fonction de voisinage h(t).

[0086] Le paramètre $\alpha(t)$ règle la vitesse de l'apprentissage et est initialisé avec une grande valeur au début puis décroît avec le nombre d'itérations en vue de ralentir au fur et à mesure le processus d'apprentissage. Le paramètre $\alpha(t)$ prend ses valeurs entre 0 et 1. La fonction h(t) définit quant à elle l'appartenance au voisinage. Elle dépend à la fois de l'emplacement des neurones sur la carte et d'un certain rayon de voisinage. La fonction h(t) prend ses valeurs entre N/2 et 0, où N représente le nombre de neurones du plus grand côté de la carte.

[0087] Dans les premières itérations, le rayon de voisinage est avantageusement assez large pour permette de mettre à jour un grand nombre de neurones voisins du neurone BMU, mais ce rayon se rétrécit progressivement pour ne contenir que le neurone BMU et ses voisins immédiats, ou bien même le neurone BMU seulement. La règle de mise à jour des vecteurs poids W est la suivante :
$W_p(t + 1) = W_p(t) + \alpha(t)h_{cp}(t)\,[V(t) - W_p(t)]$ avec $p \in \{1, ...,P\}$
où c est le neurone BMU du vecteur d'entrée V(t) présenté au réseau à l'itération t et h la fonction de voisinage qui définit la proximité entre les neurones c et p.

[0088] Une fonction de voisinage entre le neurone vainqueur c et un neurone p de la carte, qui peut être utilisée dans le cadre de la présente invention, vaut 1 si le neurone p se trouve à l'intérieur d'un carré centré sur le neurone c et 0 dans les autres cas. Le rayon de ce carré est appelé rayon de voisinage. Il est avantageusement large au début, puis se rétrécit avec le nombre d'itérations pour contenir seulement le neurone c avec ses voisins immédiats à la fin de l'apprentissage ou même seulement le neurone c. Une autre fonction de voisinage plus flexible, qui peut être utilisée dans le cadre de la présente invention, et plus commune est la fonction gaussienne définie ci-dessous :

$$h_{cp}\big(\sigma(t)\big) = exp\left(-\frac{\|r_c - r_p\|^2}{2\sigma^2(t)}\right)$$

où r_c et r_p sont respectivement l'emplacement du neurone c et du neurone p sur la carte, et $\sigma(t)$ est le rayon du voisinage à l'itération t du processus d'apprentissage.

[0089] Avec une telle fonction de voisinage, l'amplitude de l'ajustement est graduée selon l'éloignement du neurone BMU qui réserve à lui-même l'amplitude maximale.

[0090] L'apprentissage non supervisé présenté ci-dessus résulte en une projection non linéaire de l'ensemble des vecteurs descripteurs V sur la carte C. Chaque vecteur descripteur V est alloué à son neurone vainqueur BMU. Outre la tâche de quantification, cette projection préserve la topologie des données grâce à l'utilisation de la fonction de voisinage. Deux neurones p voisins sur la carte représenteront des vecteurs descripteurs V proches dans l'espace de données.

[0091] Comme mentionné ci-dessus, il est possible au lieu d'un apprentissage dit séquentiel, d'utilisé un apprentissage en mode différé. A chaque itération t, tous les vecteurs descripteurs V sont présentés au réseau de neurones p et la mise à jour des vecteurs poids W se fait en prenant en compte tous les vecteurs descripteurs V. Chaque vecteur poids W est une moyenne pondérée des vecteurs descripteurs (Vi, i ∈ {1, ... , n}) quand le carré de la distance euclidienne est utilisée pour le calcul du neurone vainqueur, les poids correspondants étant les valeurs de la fonction de voisinage h(t)

[0092] La règle de mise à jour des vecteurs poids W est donnée par :

$$W_p(t+1) = \frac{\sum_{i=1}^{n} h_{pc_i}(t) V_i}{\sum_{i=1}^{n} h_{pc_i}(t)} \quad \text{avec} \quad p \in \{1, \dots, P\}$$

où h est la valeur de la fonction de voisinage entre le neurone vainqueur c_i du vecteur V_i et le neurone p.

[0093] La mise à jour des vecteurs poids W peut être formulée autrement en utilisant le fait que les vecteurs descripteurs V qui ont le même neurone vainqueur ont la même valeur pour la fonction de voisinage et appartiennent à la région de Voronoï dont le centre est leur neurone vainqueur :

$$W_p(t+1) = \frac{\sum_{l=1}^{K} h_{pl}(t) n_l \overline{x_l}}{\sum_{l=1}^{K} n_l h_{pl}(t)} \quad \text{avec} \quad p \in \{1, \dots, P\}$$

où n_l est le nombre d'observations appartenant à la région de Voronoï représentée par le neurone l. et V_l est la moyenne des observations de cette même région.

[0094] Vers la fin de l'apprentissage, quand le rayon de voisinage devient trop petit pour activer seulement le neurone vainqueur, chaque vecteur poids W constitue le centre de gravité des vecteurs descripteurs V qu'il représente et on retombe alors sur l'algorithme des centres-mobiles, ce qui garantit une meilleure approximation de la fonction de densité des observations. De plus, avec l'absence du pas d'apprentissage α(t), cet algorithme ne présente pas de problèmes de convergence.

[0095] Cependant, le mode différé pourrait causer des torsions dans les cartes à grandes dimensions. Pour cette raison, on procède à une analyse en composantes principales pour initialiser les vecteurs prototypes.

[0096] La carte auto-organisée C peut être une carte bidimensionnelle ou tridimensionnelle. L'initialisation W de la carte auto-organisée C avant la procédure d'apprentissage en tant que telle peut être effectuée de plusieurs façons. Par exemple, une première méthode d'initialisation consiste à assigner un vecteur poids W initial à chaque nœud de la carte auto-organisée C. Cette d'attribution initiale des vecteurs poids W peut être par exemple une attribution aléatoire d'un nombre à chaque composante des vecteurs poids, sans stimulation. Le terme „aléatoire“ désigne probabilité égale pour n'importe lequel d'un ensemble de résultats possibles. La valeur numérique de ces composantes assignées au hasard peut être approximativement limitée à la borne inférieure et supérieure par l'extrema correspondant observé dans les vecteurs descripteurs, c'est-à-dire les vecteurs V. Une autre méthode d'initialisation des vecteurs poids W comprend une variation systématique, par exemple une variation linéaire, dans la plage de chaque dimension de chaque vecteur poids W pour recouper approximativement la plage correspondante observée dans les vecteurs descripteurs V. Dans une autre méthode d'initialisation, les vecteurs poids W sont initialisés par les valeurs des vecteurs ordonnés le long d'un sous-espace bidimensionnel traversé par les deux vecteurs propres principaux des vecteurs descripteurs V obtenus par des méthodes d'orthogonalisation bien connues dans l'art, par exemple par l'orthogonalisation dite de „Gram-Schmidt“. Dans une autre procédure d'initialisation, les valeurs initiales des composantes des vecteurs poids W sont fixées sur des échantillons choisis au hasard parmi les vecteurs descripteurs V.

[0097] La détermination du neurone BMU de la carte auto-organisée C pour chaque vecteur descripteur V peut se faire selon plusieurs critères bien connus de l'homme du métier. Cela peut, par exemple, être effectué sur la base d'une distance, par exemple la distance Euclidienne minimale entre tous les vecteurs poids W de la carte auto-organisée C et le vecteur descripteur V. D'autres méthodes peuvent être employées pour la détermination du neurone BMU telles que celles utilisant la corrélation entre vecteurs qui présente l'avantage d'offrir plus de robustesse au décalage entre vecteurs, l'écart angulaire entre vecteurs qui offre l'avantage de mettre l'accent sur la longueur mutuelle des vecteurs pour autant que l'information soit portée par ces grandeurs, la mesure de distance de Minkowsky qui est une généralisation de la mesure de distance euclidienne et qui est avantageuse lorsque les vecteurs portent des données de nature qualitatives peuvent être aussi mises en oeuvre.

[0098] Dans le cadre de la présente invention, les vecteurs descripteurs V sont stockés dans la base de données „vecteurs descripteurs“ 140 et le nombre de neurones p varie selon le nombre de documents de l'ensemble de documents E afin d'assurer une répartition aussi uniforme que possible des documents sur la carte auto-organisée C.

[0099] A l'issue de la phase d'apprentissage telle qu'elle a été décrite, une matrice de poids M de nombres réels est délivrée et stockée dans la base de données „matrice de poids“ 310 (voir Figure 4), dont le nombre de lignes est égal au nombre de composantes des vecteurs descripteurs de documents V de la base de données „documents normalisés“ 140 et le nombre de colonnes est égal au nombre de neurones p de la carte auto-organisée C.

[0100] Cette matrice de poids M peut être avantageusement utilisée par le processeur de génération et de traitement de la carte des documents 350 pour produire le contenu de trois bases de données „heatmap“ 320, „wordmap“ 330 et „pointmap“ 340 qui pourront être exploitées par l'intermédiaire de l'interface graphique 190 (voir ci-dessous).

[0101] La base de données „pointmap“ 340 a pour but de permettre l'affichage des documents sur une représentation graphique de la carte C. Le processeur de traitement 350 va faire appel à la base de données „vecteurs descripteurs“ 175 contenant les vecteurs descripteurs V de tous les documents de l'ensemble de documents E. Pour chaque ligne de la base de données 175, un calcul de distance va être réalisé en le vecteur descripteur V avec tous les vecteurs poids W des neurones p de la carte.

$$d(W_c, X) = \min_{p\in\{1,\ldots,P\}} d\left(W_p, X\right)$$

[0102] L'indice, ou le numéro, du neurone c dont le vecteur poids W possède la plus petite distance avec le vecteur descripteur V de document présenté est associée à ce vecteur descripteur V. Ainsi, lorsqu'il s'agira d'afficher le document dont le contenu correspond le plus à un terme de recherche R (voir ci-dessous) sur une représentation graphique de la carte C, l'indice du neurone qui répond le mieux à ce document sera disponible sans avoir à effectuer de nouveau ce calcul.

[0103] En passant en revue l'ensemble des vecteurs descripteurs V à travers le processeur de traitement 350 on obtient la base de données „pointmap“ 340 contenant tous les vecteurs descripteurs V et les indices des neurones de la carte auto-organisée C qui répondent le mieux aux vecteurs V.

[0104] A l'issu d'une recherche, le nombre de documents à afficher sur la représentation graphique de la carte C peut être très élevé et conduire à un affichage illisible si ces derniers étaient tous placés sur un même plan. Dans le cadre de cette invention, l'affichage des documents sur la carte peut être avantageusement allégé en offrant un système de zoom comparable à celui qui est disponible pour les cartes routières. Chaque indice (x,y) contenu dans la base de données „pointmap“ est enrichi par une valeur z correspondant à un plan d'affichage. Pour chaque document, la valeur z est calculée suivant une fonction d'évaluation dont le libre choix est avantageusement laissé à l'utilisateur. Ce peut être, par exemple et de manière non limitative, une évaluation manuelle ou le résultat d'un calcul. Cette fonction est déterminée en tant que paramètre du processeur de génération et de traitement de la carte de documents 350. A l'issue de l'exécution du processeur 350, la base de données „pointmap“ 340 contient tous les mots du vocabulaire ainsi que les coordonnées (x,y,z) de la carte auxquelles ces mots doivent être affichés.

[0105] Par exemple, les informations relatives au document de l'exemple concret utilisé ici sont disponibles dans la base de données „pointmap“ 340 et sont illustrées dans le Tableau 5. Un exemple d'affichage des documents est présenté sur la Figure 5.

ID	Contenu du document	Numéro de neurone	Position x	Position y	Position z
19	a vengeful New York transit cop décidé to steal a trainload of subway fare -PRON-fos-ter brother a fellow cop try to protect - PRON-	8299	99.114	82.886	0.00159114 742

Tableau 5: extrait de la base de données „pointmap“ 340

[0106] Comme mentionné ci-dessus, le but de la présente invention est de permettre d'identifier à l'aide d'une recherche analogique un ou plusieurs documents dont le contenu est le plus proche d'un terme de recherche R. Pour ce faire, un terme de recherche R, qui peut être avantageusement être saisi par un utilisateur par l'intermédiaire de l'interface graphique 190, est transformé en un vecteur de recherche K qui peut ensuite être comparé aux colonnes de la matrice de poids M de la base de données „matrice de poids“ 310. Le processus de transformation est illustré dans la **Fehler! Verweisquelle konnte nicht gefunden werden**. 6.

[0107] Chaque mot du terme de recherche R est lu séquentiellement à l'étape 910 puis il est extrait à l'étape 920 pour être comparé à la liste des mots qui composent le vocabulaire 160. Si le mot lu est un mot du vocabulaire, l'indice auquel ce mot se trouve dans la base de données „vocabulaire“ 160 est enregistré à l'étape 940 et le processus de lecteur et de comparaison se poursuit jusqu'à ce que tous les mots du champ de recherche aient été lus et que la lecture des mots soit terminées à l'étape 950. A l'issue de ce processus, une liste d'indices 960 qui correspondent aux mots du terme de recherche R qui ont été trouvés dans la base de données „vocabulaire“ 160 est obtenue. Avantageusement, la valeur 1 est stockée à l'emplacement de ces indices pour former un vecteur de recherche K dont le nombre de composantes est

égal au nombre de mots identifiés. Le vecteur K est finalement normalisé en utilisant avantageusement une normalisation de type „L2".

[0108] Une distance est ensuite calculée entre les valeurs qui se trouvent enregistrées dans ces indices aux valeurs enregistrées dans la base de données „matrice de poids" 310 qui se trouvent aux mêmes indices 960. En d'autres termes, une distance est calculée entre le vecteur de recherche K et tous les vecteurs poids W de la carte auto-organisée C. Le ou les neurones pbest de la carte auto-organisée C qui répondent le plus, c'est-à-dire ceux pour lesquels la distance calculée est la plus petite, sont identifiés sur la carte. Les identifiants des documents qui sont rattachés à ces neurones sont extraits en utilisant la base de données „pointmap" 340 et les signets sont avantageusement affichés sur une représentation graphique d'une carte de documents CD qui peut être superposée sur la carte auto-organisée C, comme cela est montré sur la **Fehler! Verweisquelle konnte nicht gefunden werden**..

[0109] Il est bien entendu possible que la liste des documents identifiés sur la base d'une recherche analogique et du terme de recherche R soit mise à disposition d'une autre manière que sur une carte de documents CD. Par exemple, une simple liste des ID des documents identifiés pourrait représenter le résultat de la recherche analogique. Néanmoins, il est important de souligner qu'indépendamment de la représentation graphique des documents trouvés, ces documents sont toujours identifiés sur la base de la distance entre le vecteur de recherche K et les vecteurs de poids W de la carte auto-organisée. Le vecteur poids W dont la distance D est minimale avec le vecteur de recherche K est dans un premier lieu identifié. Comme expliqué ci-dessus, chaque vecteur poids W correspond à un neurone p de la carte auto-organisée C. De plus, lors de la génération de la carte auto-organisée C, chaque document de l'ensemble de document E a été alloué à un neurone de la carte C. Ceci permet donc en connaissant quel vecteur poids W est le plus proche du terme de recherche R de déterminer quel document est alloué au neurone correspondant et donc quel document a le contenu le plus proche du terme de recherche R.

[0110] Comme mentionné auparavant, l'algorithme SOM a pour effet de regrouper les documents proches au sens d'une mesure de distance dans des neurones. Les neurones sont codés dans une matrice sous forme de vecteurs de données réelles. Ces neurones sont ordonnés par l'algorithme de telle sorte que des documents proches dans l'espace des données soient aussi proches que possibles sur la carte C. Il s'agit d'une des propriétés les plus importantes de cet algorithme. Néanmoins, à cause de la forte non-linéarité de la projection des données d'origine, c'est-à-dire les vecteurs descripteurs V, sur la carte C, la proximité entre les neurones ne donne aucune information sur la distance réelle qui sépare les vecteurs descripteurs V dans l'espace d'origine. Il en résulte que des documents alloués à des neurones proches sur la carte C peuvent en réalité correspondre à des données très distantes dans l'espace d'origine et donc en réalité être très différents. Cette limitation peut être en partie réduite par l'utilisation d'une carte de distances CH qui peut être superposée à la carte auto-organisée C dans l'interface graphique 190. Cette carte de distances CH est une carte de dimension égale à la carte C et qui donne une mesure de la distance réelle entre les vecteurs poids W de cette dernière. Cette mesure peut être avantageusement affichée sur une représentation graphique de la carte de distances CH par une „lookup table" adaptée, par exemple une coloration adaptée.

[0111] Dans le cadre de cette invention, chaque point de la carte de distances CH, aussi appelée „heatmap", est associé à une valeur dd_{ij} calculée de la manière suivante :

$$dd_{ij} = \sum_{k=i-1}^{i+1} \sum_{l=j-1}^{j+1} d(W_{ij}, W_{k,l})$$

avec k, l ∈ {-1,0, +1} et d est la mesure de distance euclidienne

[0112] A l'issue de l'exécution du processeur de traitement 186, la base de données „heatmap" 320 contient les coordonnées de chaque point ij de la carte de distance CH ainsi que la valeur dd_{ij} calculée pour ce point. Avantageusement, l'échelle de couleur pour représenter ces valeurs peut s'étendre du rouge au vert où le rouge représente la valeur la plus élevée et le vert la distance la plus éloignée. Un exemple de carte de distances CH avec une grille de 10'000 neurones est montré sur la Fehler! Verweisquelle konnte nicht gefunden werden.. Dans cette figure le code couleur est représenté par des niveaux de gris.

[0113] Tout comme une carte routière représente les villes dans leur contexte incluant les monuments, les routes, les forêts et en général tout ce qui donne un contexte à la ville, la représentation des documents sur une carte peut être contextualisée en plaçant sur une carte de mots CW les mots les plus significatifs à proximité des neurones de la carte C correspondant aux documents qui contiennent ces mots. La carte de mots CW peut être sur l'interface graphique 190 avantageusement superposée à la carte auto-organisée C afin de donner une information supplémentaire à l'utilisateur.

[0114] Dans une version standard le l'algorithme des cartes auto-organisées, le positionnement des mots d'un document sur la carte de mots CW à l'emplacement du neurone de la carte C correspondant est possible mais tous les mots sont ainsi placés au même endroit, à savoir la position du neurone sur la carte C. La représentation graphique de la carte de mots CW est alors inutilisable lorsque le nombre de mots devient élevé car ils sont tous superposés au même emplacement. L'une des originalités apportées par la présente invention est d'offrir une représentation nouvelle de l'ensemble des documents traités sous forme d'une carte. Les mots les plus significatifs vont être placés de manière continue sur la carte de mots

CW selon la méthode présentée ici. Rappelons tout d'abord que les neurones sont ordonnés suivant une relation d'ordre mono-, bi- ou tridimensionnelle selon l'application visée. Dans le cadre de l'exemple d'application de la présente invention, nous considérons une relation d'ordre bidimensionnelle.

[0115] Pour produire les mots les plus significatifs rattachés à un neurone, le processeur de traitement 350 va identifier pour chaque neurone les indices du vecteur poids W correspondant dont la composante dépasse un seuil prédéterminé. Il va ensuite extraire de la base de données „vocabulaire" 160 les mots qui sont situés aux mêmes indices pour les rattacher au neurone considéré. Le principe est illustré dans la Figure 8 où le point (2,3) de la carte de mots CW se verra rattacher les mots „James", „Spy" et „Bridge". De plus, pour chaque mot la valeur de la composante du vecteur poids qui lui correspond est enregistrée. Ce processus de rattachement est réalisé pour l'ensemble des points de la carte de mots CW comme illustré sur la Figure 9a. Ensuite, pour chaque mot, une liste des points de la carte de mots CW auxquels il est rattaché est établie. Un exemple d'une telle liste se trouve dans la Figue 9b. Avec cette liste et la valeur de la composante du vecteur poids W pour l'indice du mot choisi, le processeur de traitement 350 va calculer un indice continu des emplacements en multipliant la valeur des coordonnées des neurones de la liste par la valeur de la composante du vecteur poids W(troisième colonne de la Figure 9b)

[0116] A l'issue de cette étape de calculs, une liste de valeurs réelles qui peuvent utilisées pour calculer la position du mot sur la carte est disponible. A cet effet, un calcul de barycentre va être utilisé. La somme des valeurs de la composante du vecteur poids W pour chaque mot choisi est calculée :

$$Wtotal_m = \sum_{(k,l)} w_{m,(k,l)}$$

avec (k, l) désignant les indices des neurones pour lesquels $w_{i,(k,l)}$ > seuil prédéterminé et m désigne l'indice du vecteur poids pour le mot choisi.

[0117] Dans l'exemple des Figures 9a et 9b, la valeur Wtotal pour le mot „spy" vaut :
$Wtotal_{spy}$ = 0.1 + 0.3 + 0.1 + 0.2 + 0.74 + 0.1 + 0.1 + 0.1 = 1.74

[0118] Enfin, la somme des coordonnées de l'indice continu des emplacements est calculée pour être divisée par Wtotal du mot choisi. Dans l'exemple des Figures 9a et 9b, la somme des coordonnées vaut :
(0.1,0,4) + (0.6,1.2) + (0.3,0.4) + (0.2,0.6) + (1.48,2.22) + (0.3,0.3) + (0.2,0.2) + (0.3,0.2) = (3.48,5.52)

[0119] La division de chacune des coordonnées par Wtotal du mot donne finalement (2,3.17). Le mot „spy" serait donc placé sur la représentation graphique de la carte de mots CW à l'emplacement (2,3.17) comme cela est illustré dans la Figure 10.

[0120] Selon la taille du vocabulaire, le nombre de mots à afficher sur la carte peut être très grand. Si tous les mots étaient placés sur un même plan, cela pourrait conduire à un affichage illisible des mots sur la carte.

[0121] Dans le cadre de cette invention, il a été prévu, tout comme pour l'affichage des documents sur la représentation graphique de la carte de documents CD, d'alléger l'affichage des mots sur la carte de mots CWen offrant un système de zoom comparable à celui qui est disponible pour les cartes routières. Chaque position (x,y) d'un mot sur la carte de mots CW va être enrichie par une valeur z utilisée comme plan d'affichage, comme cela est montré sur la Figure 11. Sur cette figure, les mots sont placés sur 2 plans d'affichage de la carte de mots CW suivant les valeurs z qui leur sont associées.

[0122] Pour chaque mot, la valeur z est calculée suivant une fonction d'évaluation dont le libre choix est laissé à l'utilisateur. Ce peut être, par exemple et de manière non limitative, une évaluation manuelle ou le résultat d'un calcul. Cette fonction est déterminée en tant que paramètre du processeur de génération et de traitement de la carte de documents 350. A l'issue de l'exécution du processeur 350, la base de données „wordmap" 330 contient tous les mots du vocabulaire ainsi que les coordonnées (x,y,z) de la carte de mots CW auxquelles ces mots doivent être affichés. Un extrait de la base de données „wordmap" 330 est présenté dans le Tableau 6.

Mot	x	y	z
revenge	2.95	97.96	0.01
revenge	7.00	52.00	0.01
revenge	11.00	92.00	0.01
revenge	13.99	61.98	0.01
revenge	16.31	47.52	0.01
revenge	20.86	39.70	0.01

Mot	x	y	z
revenge	23.00	28.00	0.01
revenge	25.02	91.71	0.01
revenge	30.32	53.95	0.01
revenge	31.16	73.14	0.01
revenge	40.02	99.00	0.01
revenge	51.00	47.52	0.01
revenge	51.00	33.00	0.01
revenge	51.49	73.00	0.01
revenge	62.75	96.49	0.07
revenge	64.00	47.00	0.01
revenge	83.31	53.73	0.01
revenge	84.67	93.33	0.01
revenge	90.99	63.51	0.01
revenge	94.70	94.85	0.01
revenge	98.14	75.86	0.01

Tableau 6: extrait de la base de données „wordmap“ 330 pour le mot „revenge“

[0123] Comme on peut le constater dans ce tableau, le mot „revenge“ est identifié à 21 emplacements différents sur la carte de mots CW qui correspondent à autant de contextes différents. De plus, ce même mot a une importance plus élevée (valeur z) à la position (62.75, 96.49). Il pourra ainsi être placé sur un plan d'affichage de l'interface graphique 190 plus élevé et sera donc mis en évidence avec plus d'importance à cet emplacement.

[0124] Par cette méthode, tous les mots du vocabulaire peuvent être placés sur la carte de mots CW et superposés à la carte auto-organisée C définissant ainsi un contexte lexical de cette dernière. Un exemple est montré sur la **Fehler! Verweisquelle konnte nicht gefunden werden**. 12.

[0125] Comme mentionné ci-dessus, la présente invention prévoit avantageusement une interface graphique 190 pour permette une représentation graphique de la carte auto-organisée C, de la carte de mots CW, de la carte de distance CH ainsi que la carte de documents CD. Un mode de réalisation de cette interface graphique 190 est présenté dans la **Fehler! Verweisquelle konnte nicht gefunden werden**.. La zone d'affichage 485 de l'interface graphique 190 est destinée à présenter une représentation graphique de ces cartes. Les données requises pour cet affichage sont stockées dans les bases de données 310, 320, 330 et 340.

[0126] Le bouton 470 permet à l'utilisateur d'effectuer un zoom avant avec le symbole + ou un zoom arrière avec le symbole - sur la carte. Cela a pour effet d'agrandir ou de diminuer la zone allouée à chaque neurone pour permettre un affichage de son contenu plus détaillé. Le bouton 480 permet d'afficher ou de masquer les emplacements des neurones sous forme de grille. Le bouton 490 permet à l'utilisateur de revenir à l'affichage d'origine de la carte. Le bouton 460 permet à l'utilisateur de montrer ou de cacher la carte de mots CW. Le bouton 450 permet à l'utilisateur de montrer ou cacher la carte de document CD dans laquelle les documents sont représentés par des épingles. Le bouton 440 permet à l'utilisateur de montrer ou cacher la carte de distances CH. Le champ de saisie 400 permet à l'utilisateur d'écrire un terme de recherche R.

[0127] Selon le choix effectué par l'utilisateur avec le bouton 410, la recherche s'effectuera suivant un type „analogique“ ou „textuel“. Pour ce qui est du mode de recherche „textuel“ voir ci-après.

[0128] La zone d'information 430 donne à l'utilisateur des informations sur le nombre de documents utilisés pour construire la carte, le nombre de points affichés dans la vue choisie par l'utilisateur et la taille du vocabulaire.

[0129] La liste déroulante 420 permet à l'utilisateur de choisir une base de données de documents dans l'ensemble des bases de données disponibles.

[0130] La zone d'affichage 495 permet à l'utilisateur d'afficher la totalité d'un document sélectionné sur la carte. Lorsque l'utilisateur clique avec le bouton gauche d'une souris d'ordinateur sur une épingle représentant un document sur la carte

de documents CD, une fenêtre „popup“ de résumé 475 est affichée. Lorsque l'utilisateur clique avec le bouton gauche sur le lien „Show all“ affiché dans la fenêtre popup de résumé, l'intégralité du document est affichée dans la zone d'affichage 495.

[0131] Avantageusement, l'interface graphique 190 est configurée pour qu'initialement la carte auto-organisée C soit affichée et en superposition la carte de distance CH, la carte de mots CW et la carte de documents CD, comme cela est montré sur la Figue 14.

[0132] L'utilisateur peut faire appel aux différents contrôles disponibles pour explorer le contenu de l'affichage, réaliser des recherches et affiner les résultats obtenus. L'utilisateur dispose par exemple de deux boutons 470 pour zoomer vers l'avant ou vers l'arrière sur une partie des cartes. Lorsqu'il clique sur le zoom (avant ou arrière), l'événement est capté par l'interface puis transmis au processeur de génération et de de traitement 350. Il détermine la zone actuellement en cours d'affichage, fait appel aux bases de données „wordmap“ 330 et „pointmap“ 340 pour rechercher les mots et les documents qui sont dans la zone identifiée. Il sélectionne les mots et les documents en fonction de la valeur z qui leur est associée et renvoie la liste au processeur qui est en charge de l'affichage.

[0133] L'utilisateur dispose également avantageusement du bouton 480 pour afficher ou masquer l'emplacement des neurones sur la carte. Cet affichage permet d'identifier quels sont les documents rattachés à chaque neurone. L'affichage de la taille de la grille dépend du niveau de zoom sur la carte.

[0134] L'utilisateur dispose de 3 boutons indépendants :

- Show/hide words 460 qui permet d'afficher ou de masquer les mots sur la carte de mots CW, quel que soit le niveau de zoom courant ;
- Show/hide points 450 qui permet d'afficher ou de masquer les pointeurs des documents sur la carte de documents CD ; et
- Show/hide heatmap 440 qui permet d'afficher ou de masque la carte de distances CH.

[0135] Chaque pointeur de document représenté sur la carte de documents CD par un signet rouge est un élément cliquable avec le bouton droit de la souris. Lorsque le signet est cliqué, une fenêtre d'affichage 475 est montrée à l'utilisateur comme cela est représenté sur la Figure 15. Elle contient un résumé du document sélectionné ainsi qu'un lien „Show all“ vers le contenu complet du document.

[0136] Lorsque l'utilisateur clique sur le lien „Show all“, une fenêtre additionnelle 495 est affichée en partie droite de l'interface 190 avec la totalité du document comme montré sur la Figure 16.

[0137] L'utilisateur peut alors cliquer sur la croix de fermeture pour fermer chaque fenêtre d'affichage. Il peut cliquer sur plusieurs signets pour afficher plusieurs fenêtres de résumé mais seule la dernière fenêtre d'affichage de document complet peut être affichée.

[0138] Le mode de recherche analogique présenté plus haut et permet de rechercher le terme de recherche R saisi dans le champ de saisie 400 suivant un mode de calcul qui exploite les valeurs enregistrées la base de données „matrice de poids“ 310. Nous rappellerons simplement ici, que le ou les neurones qui répondent le plus, c'est-à-dire ceux pour lesquels la distance calculée entre un vecteur de recherche K du terme de recherche R et les vecteurs poids de la carte auto-organisée C est la plus petite, sont identifiés sur la carte. Les identifiants des documents qui sont rattachés à ces neurones sont extraits en utilisant la base de données „pointmap“ 340 et des signets sont affichés sur la carte de documents CD, comme cela est montré sur les **Fehler! Verweisquelle konnte nicht gefunden werden**.

[0139] Si l'utilisateur a accès à plusieurs sources de données différentes, il peut choisir celle qu'il peut explorer grâce à la liste déroulante 420 qui affiche toutes les sources de données disponibles. La sélection d'une source de données réinitialise l'affichage de la carte et prend en compte les paramètres propres à cette source de données et notamment le nombre de neurones, le vocabulaire et le nombre total de documents.

[0140] Le mode de réalisation préféré de la présente invention prévoit en plus du mode de recherche analogique un mode de recherche textuel. Ce dernier emploie un système d'indexation et de recherche textuelle de documents qui a pour objectif de créer une base de données d'index des contenus de documents afin de l'utiliser pour permettre une recherche textuelle dans le contenu des documents, avantageusement à travers l'interface graphique 190.

[0141] Le système d'indexation et de recherche textuelle est basé sur toute solution permettant d'indexer des documents. Le système d'indexation et de recherche textuelle de documents 125 représenté sur la **Fehler! Verweisquelle konnte nicht gefunden werden**., construit après la phase d'indexation une base de données d'index 220 qui pourra être utilisée par l'interface d'utilisation de la carte de documents 190 représentée sur la Figure 13.

[0142] Le champ de recherche 400 est utilisé pour entrer un terme de recherche à rechercher dans l'ensemble des documents indexés, pour autant que l'utilisateur ait fait le choix de la recherche textuelle grâce au bouton 410 de l'interface 190, positionné en mode „search type : text“. Le texte entré dans ce champ est transmis au système de recherche via l'interface de programmation API 230 qui transmet au processeur d'indexation et de recherche. Le processeur d'indexation et de recherche 230 délivre en sortie via l'interface API la liste 240 des identifiants des documents dans lesquels tous les mots sont présents. Cette liste d'identifiants est transmise au processeur de génération et de traitement 180 de la carte de documents qui va croiser la liste des identifiants avec les données de la base de données „pointmap“ 340 afin de déterminer les positions (x,y,z) de chaque document pour les placer sur la carte de documents CD.

[0143] Le résultat visuel sera identique à celui de la recherche analogique tel que présenté sur la **Fehler! Verweisquelle konnte nicht gefunden werden**.. Néanmoins, dans le cas de la recherche textuelle, tous les mots trouvés par le moteur de recherche 125 seront avantageusement surlignés en vert ou bleu dans le résumé ou dans le texte complet du document, comme cela est montré sur la **Fehler! Verweisquelle konnte nicht gefunden werden**..

[0144] Lorsque le moteur de recherche 125 délivre un résultat exact pour les mots recherchés, ces dernier sont avantageusement surlignés en vert. Si le résultat est approchant, les mots sont surlignés en bleu.

[0145] Avantageusement, la présente méthode est mise en oeuvre en utilisant un programme informatique pour effectuer des opérations sur des aspects de la présente invention qui peut être écrit dans n'importe quelle combinaison d'un ou plusieurs langages de programmation, y compris un langage de programmation orienté objet tel que Java, Python, C++, ou similaire et des langages de programmation procédurale classiques, tels que le langage de programmation „C“ ou des langages de programmation similaires. Le code du programme peut être exécuté entièrement sur l'ordinateur de l'utilisateur, en partie sur l'ordinateur de l'utilisateur, en tant que progiciel autonome, en partie sur l'ordinateur de l'utilisateur et en partie sur un ordinateur distant ou entièrement sur l'ordinateur distant ou un serveur. Dans ce dernier scénario, l'ordinateur distant peut être connecté à l'ordinateur de l'utilisateur via n'importe quel type de réseau, y compris un réseau local (LAN) ou un réseau étendu (WAN), ou la connexion peut être établie à un ordinateur externe (par exemple, par le biais de la fonction Internet à l'aide d'un fournisseur de services Internet).

[0146] L'ordinateur exécutant le programme sera composé au minimum d'un processeur standard (CPU) avec sa mémoire RAM d'au minimum 30Giga octets, un disque dur de capacité minimum 1Tera Octet. Il pourra être aussi composé d'un processeur d'exécuter plusieurs threads simultanément (multi-thread). Enfin, il peut lui être adjoint des cartes d'accélération matérielles telles que les GPU (graphie processor Units), les TPU (Tensor Processing Units) et en général tout dispositif d'accélération matérielle disponible sur le marché tels que RTX2060, RTX 2070, GTX 1070.

[0147] Il est évident que la présente invention est sujette à de nombreuses variations quant à sa mise en oeuvre. Bien qu'un mode de réalisation non limitatif ait été décrit à titre d'exemple, on comprend bien qu'il n'est pas concevable d'identifier de manière exhaustive toutes les variations possibles. Il est bien sûr envisageable de remplacer un moyen décrit par un moyen équivalent sans sortir du cadre de la présente invention. Toutes ces modifications font partie des connaissances communes d'un homme du métier dans le domaine technique de la présente invention.

Revendications

1. Méthode mise en oeuvre par ordinateur pour la recherche analogique de documents comprenant au moins une information textuelle d'un ensemble de documents E compris dans une première base de données et dont l'information textuelle correspond le plus à un terme de recherche R comprenant les étapes :
   a. Génération d'une deuxième base de données comprenant une liste de mots produite par lemmatization de l'information textuelle des documents de la première base de données ;
   b. Génération d'un vecteur descripteur V de valeurs numériques pour chaque document de la première base de données à l'aide d'une fonction de vectorisation F de l'information textuelle ;
   c. Apprentissage d'une carte auto-organisée C comprenant un réseau de P neurones p par projection des vecteurs descripteurs V sur la carte auto-organisée C, chaque neurone p de la carte auto-organisée C correspondant à un vecteur poids W de valeurs numériques ;
   d. Allocation de chaque document de la première base de données au neurone p de la carte auto-organisée C dont le vecteur poids W correspondant possède la plus petite distance avec le vecteur descripteur V du document à allouer;
   e. Génération à l'aide de la fonction de vectorisation F et de la deuxième base de données d'un vecteur de recherche K de valeurs numériques pour le terme de recherche R ;
   f. Détermination du neurone pbest de la carte auto-organisée C dont le vecteur poids W possède la plus petite distance avec le vecteur de recherche K; et
   g. Détermination des documents de la première base de données alloués au neurone pbest de la carte auto-organisée C.
2. Méthode selon la revendication 1, dans laquelle la distance entre deux vecteurs est une distance euclidienne.
3. Méthode selon l'une des revendications précédentes, dans laquelle le contenu textuel des documents de la première base de données est normalisé avant la génération des vecteurs descripteurs V.
4. Méthode selon l'une des revendications précédentes, dans laquelle la fonction de vectorisation F est une fonction dite de „term frequency-inverse document frequency“.
5. Méthode selon l'une de revendications précédentes, dans laquelle la carte auto-organisée C est bi- ou tridimensionnelle.
6. Méthode selon l'une des revendications précédentes, dans laquelle la carte auto-organisée C est affichable graphiquement par l'intermédiaire d'une interface graphique (190) qui est par exemple l'interface graphique d'un ordinateur personnel.

7. Méthode selon la revendication 6, dans laquelle une représentation graphique d'une carte de documents CD de dimension égale à la carte auto-organisée C est superposable à la représentation graphique de la carte auto-organisée C, et dans laquelle les documents déterminés à l'étape g. sont identifiés graphiquement sur la carte de documents CD.
8. Méthode selon la revendication 7, dans laquelle le contenu d'un document déterminé à l'étape g. est accessible en sélectionnant ce document sur la représentation graphique de la carte de documents CD par l'intermédiaire d'un dispositif de pointage pour ordinateur, tel que par exemple une souris informatique.
9. Méthode selon l'une des revendications 6 à 8, dans laquelle une représentation graphique d'une carte de distances CH de dimension égale à la carte auto-organisée C est superposable à la représentation graphique de la carte auto-organisée C, dans laquelle une valeur de distance dd est attribuée à chaque neurone de la carte de distances CH, la valeur de distance dd correspondant à la somme des distances euclidienne entre le vecteur poids W du neurone considéré et les vecteurs poids W des neurones voisins directs.
10. Méthode selon la revendication 9, dans laquelle une couleur d'un codage couleur est attribuée à chaque neurone de la carte de distance CH en fonction de la valeur dd.
11. Méthode selon l'une des revendications 6 à 10, dans laquelle une représentation graphique d'une carte de mots CW de dimension égale à la carte auto-organisée C est superposable à la représentation graphique de la carte auto-organisée C, dans laquelle à chaque neurone de la représentation graphique de la carte de mots CW sont affichés les mots du vocabulaire dont la composante du vecteur poids W du neurone de la carte auto-organisé C correspondant est supérieure à une valeur prédéterminée.
12. Méthode selon la revendication 11, dans laquelle les mots affichés dans la carte de mots CW sont organisés sur des plans différents correspondant à des différentes plages de valeurs de poids, les mots dont la valeur de poids sont les plus élevées étant affichées au premier plan de la représentation graphique de la carte de mots CW.
13. Méthode selon l'une des revendications précédentes, dans laquelle les documents de la première base de données sont indexés dans une troisième base données, et dans laquelle les documents comprenant le terme de recherche R sont déterminés.
14. Méthode selon la revendication 13, dans laquelle les documents comprenant le terme de recherche R sont identifiés sur la représentation graphique de la carte de documents CD.
15. Méthode selon l'une des revendications précédentes, dans laquelle les documents de la première base de données sont des documents codés sous forme numérique avantageusement issus de traitement de texte, de systèmes de reconnaissance de textes, par exemple par l'intermédiaire de la méthode dite Optical Character Recognition, ou issus de tout système capable de produire des fichiers numériques structurés.

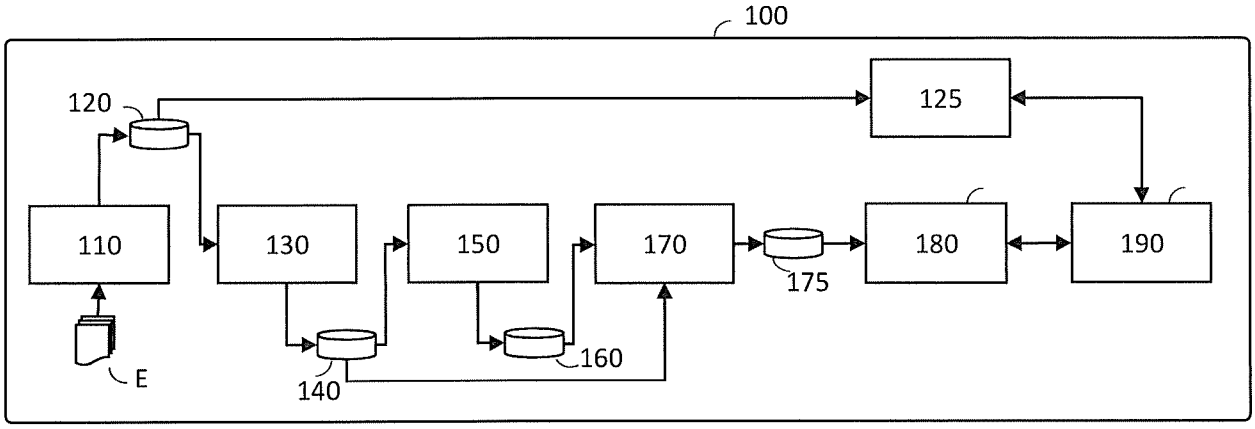

100
120
125
110
130
150
170
175
180
190
E
140
160

FIG. 1

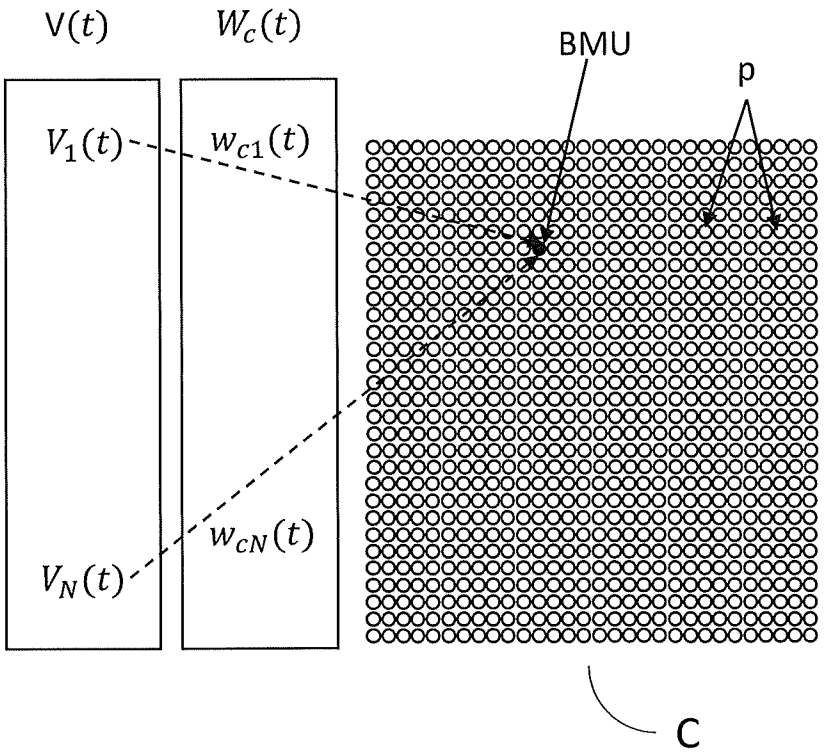

V(t)
$W_c(t)$
BMU
p
$V_1(t)$
$w_{c1}(t)$
$w_{cN}(t)$
$V_N(t)$
C

FIG. 2

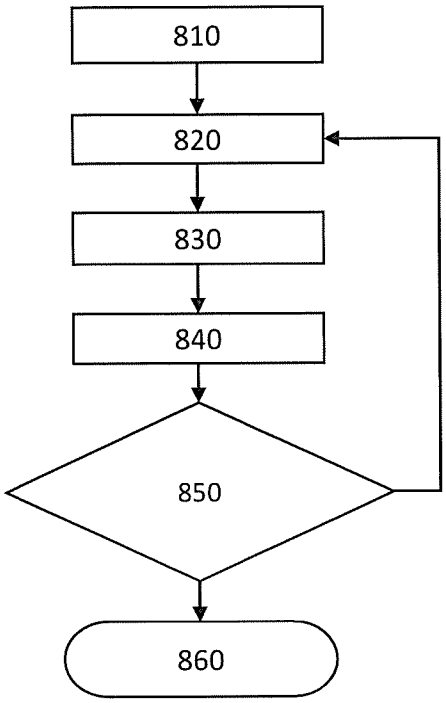


FIG. 3

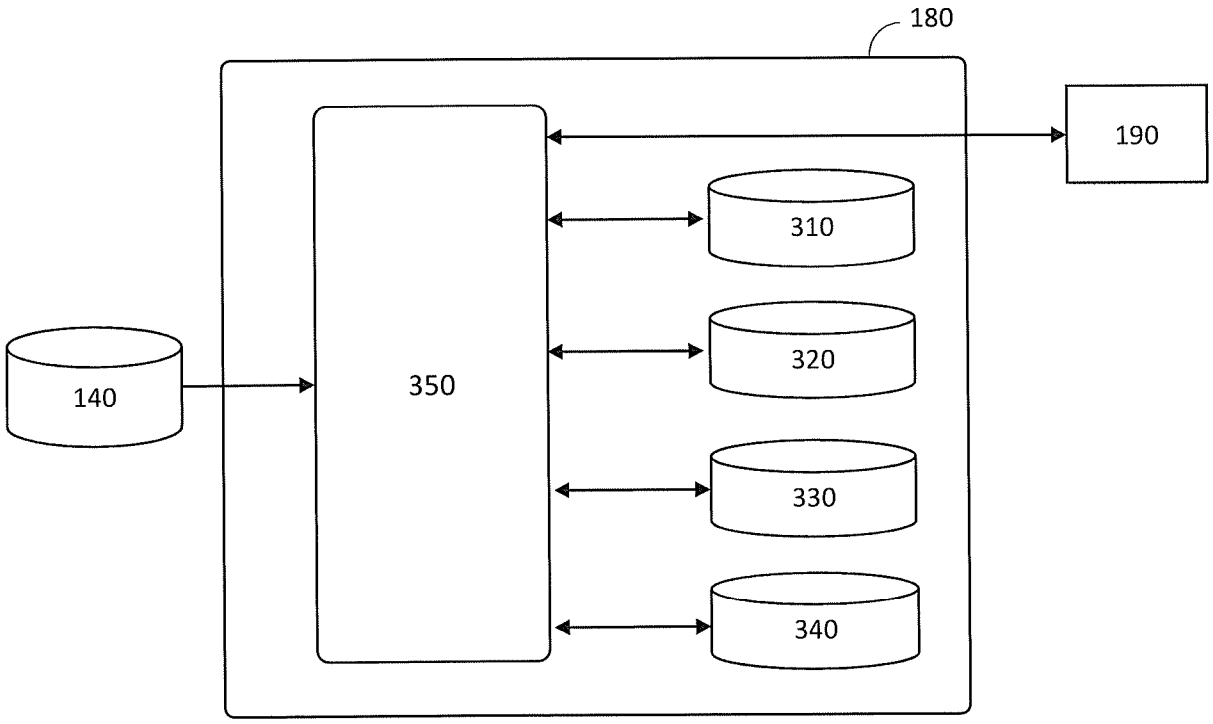


FIG. 4

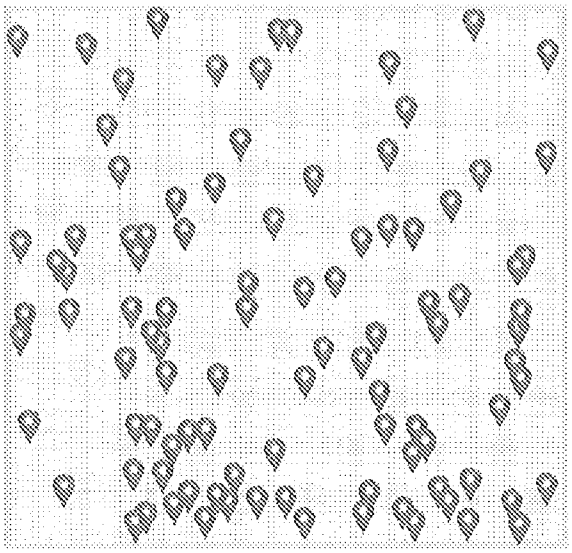


FIG. 5

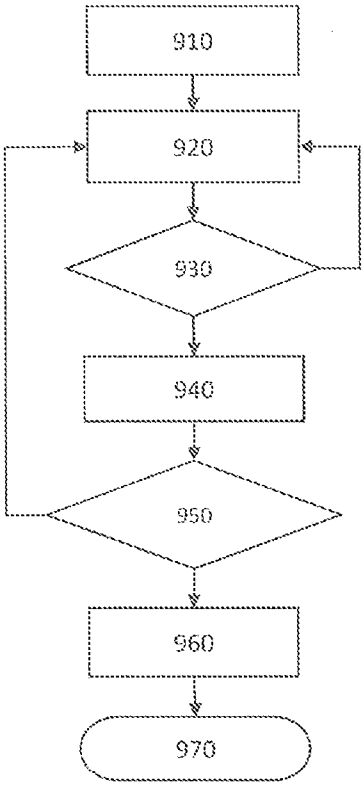


FIG. 6

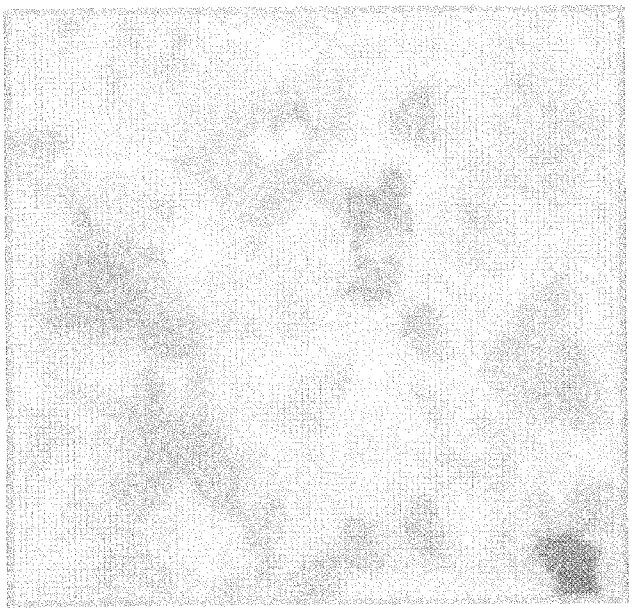


FIG. 7

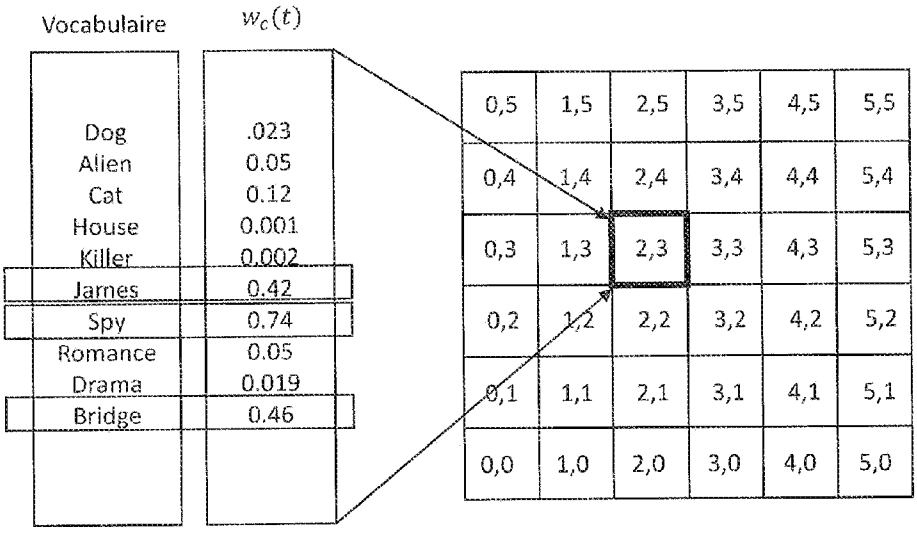

Vocabulaire
$w_c(t)$
Dog .023
Alien 0.05
Cat 0.12
House 0.001
Killer 0.002
James 0.42
Spy 0.74
Romance 0.05
Drama 0.019
Bridge 0.46
0,5 1,5 2,5 3,5 4,5 5,5
0,4 1,4 2,4 3,4 4,4 5,4
0,3 1,3 2,3 3,3 4,3 5,3
0,2 1,2 2,2 3,2 4,2 5,2
0,1 1,1 2,1 3,1 4,1 5,1
0,0 1,0 2,0 3,0 4,0 5,0

FIG. 8

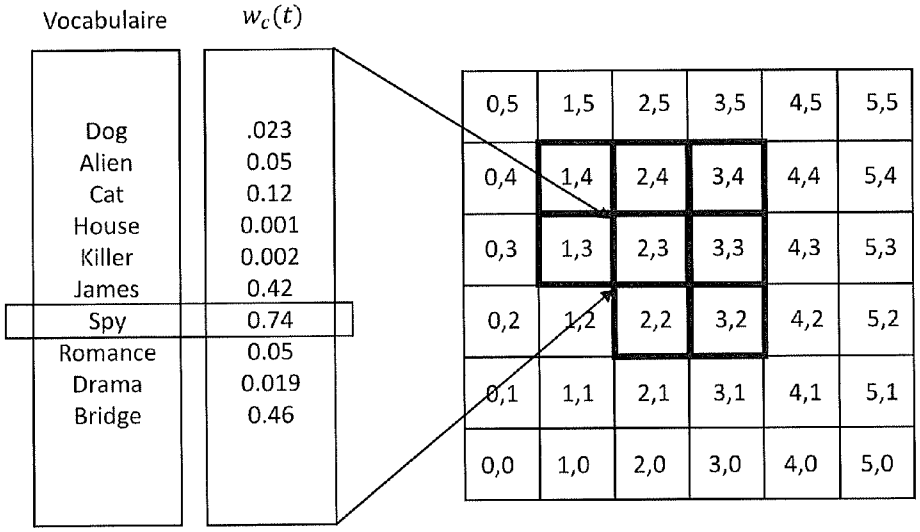
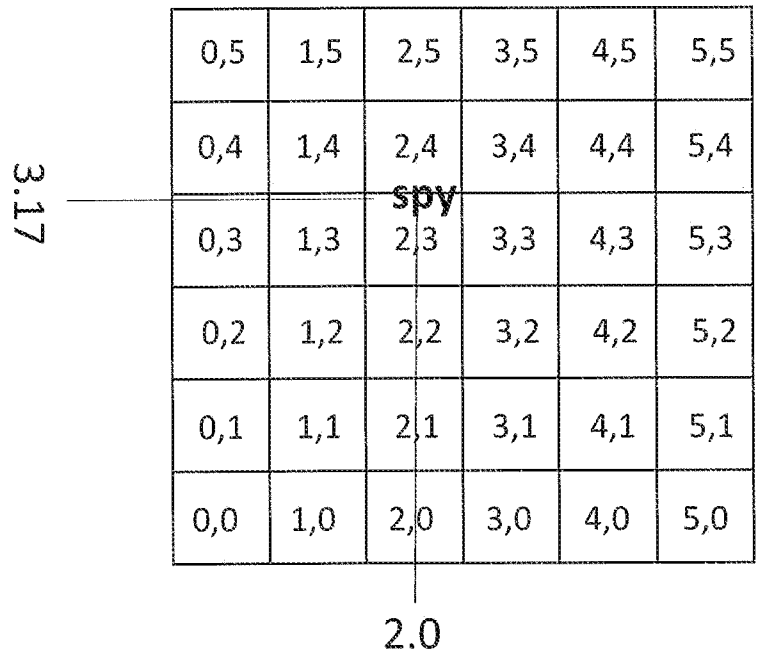


FIG. 9a

Emplacement du cluster	Valeur du poids pour l'indice du mot choisi (spy)	Index continu des emplacements
0,4	0.0	
1,4	0.1	0.1,0.4
2,4	0.3	0.6,1.2
3,4	0.1	0.3,0.4
4,4	0.0	
5,4	0.0	
0,3	0.0	
1,3	0.2	0.2,0.6
2,3	0.74	1.48,2.22
3,3	0.1	0.3,0.3
4,3	0.0	
5,3	0.0	
0,2	0.0	
1,2	0.0	
2,2	0.1	0.2,0.2
3,2	0.1	0.3,0.2
4,2	0.0	

FIG. 9b



0,5	1,5	2,5	3,5	4,5	5,5
0,4	1,4	2,4	3,4	4,4	5,4
0,3	1,3	2,3	3,3	4,3	5,3
0,2	1,2	2,2	3,2	4,2	5,2
0,1	1,1	2,1	3,1	4,1	5,1
0,0	1,0	2,0	3,0	4,0	5,0

FIG. 10

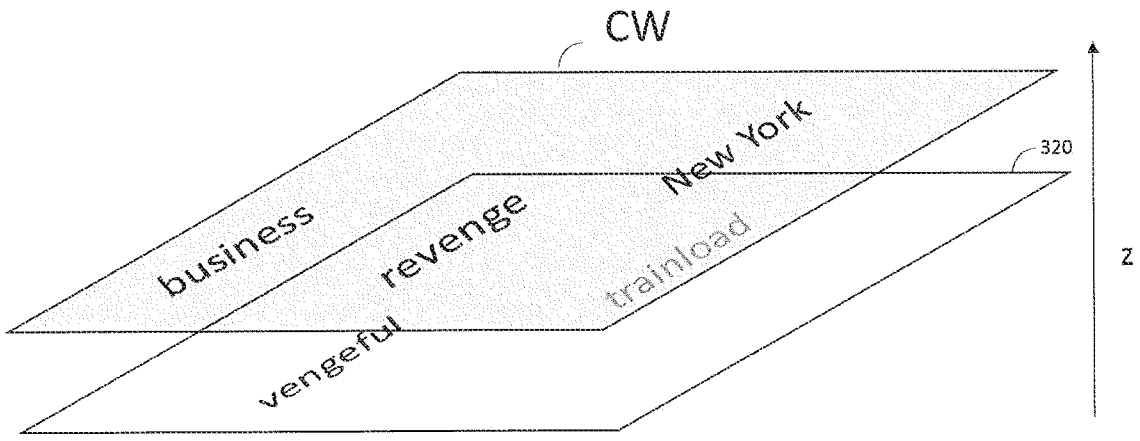


FIG. 11

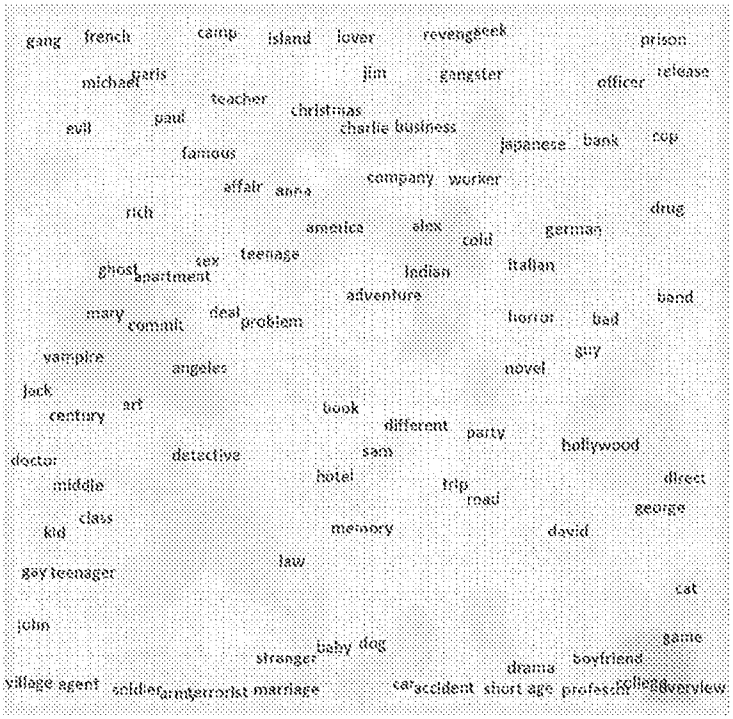

gang
french
camp
island
lover
revenge
geek
prison
michael
paris
jim
gangster
release
officer
teacher
christmas
evil
paul
charlie
business
japanese
bank
cop
famous
company
worker
affair
anna
rich
drug
america
alex
cold
german
ghost
apartment
sex
teenage
indian
italian
adventure
band
mary
commit
deal
problem
horror
bad
guy
vampire
angeles
novel
jack
century
art
book
different
party
hollywood
doctor
detective
sam
hotel
trip
road
direct
middle
george
class
kid
memory
david
law
gay teenager
cat
john
game
stranger
baby
dog
drama
boyfriend
village agent
soldier
terrorist
marriage
car accident
short age
professor
college
everyone

FIG. 12

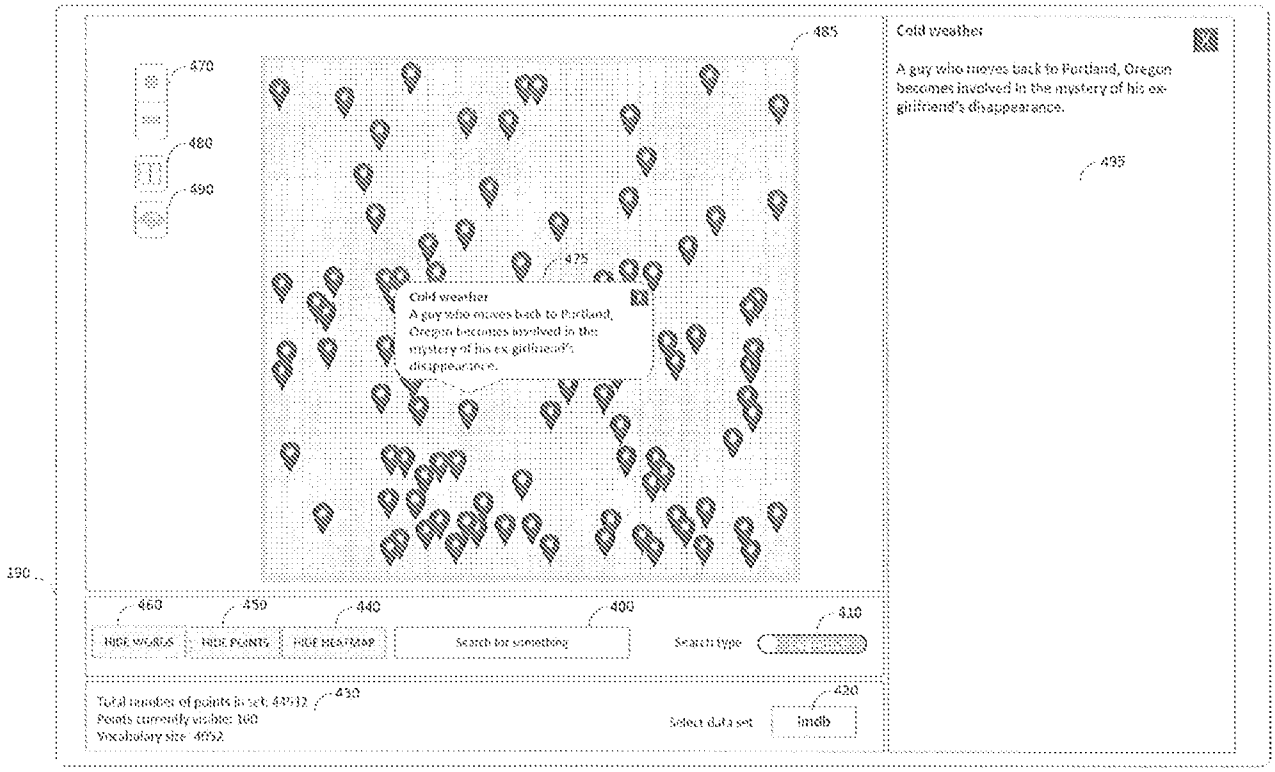

485
470
480
490
Cold weather
A guy who moves back to Portland, Oregon becomes involved in the mystery of his ex-girlfriend's disappearance.
495
475
Cold weather
A guy who moves back to Portland, Oregon becomes involved in the mystery of his ex-girlfriend's disappearance.
100
460
450
440
400
Search for something
Search type
410
Total number of points in set: 44932
430
Points currently visible: 300
Vocabulary size: 46952
Select data set
imdb
420

FIG. 13

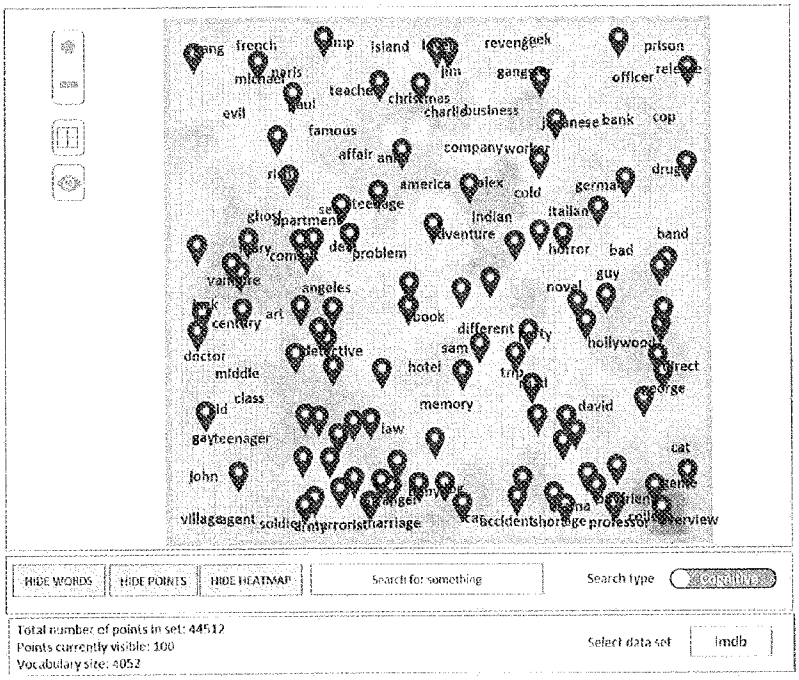

HIDE WORDS
HIDE POINTS
HIDE HEATMAP
Search for something
Search type
Total number of points in set: 44512
Points currently visible: 100
Vocabulary size: 4052
Select data set
imdb

FIG. 14

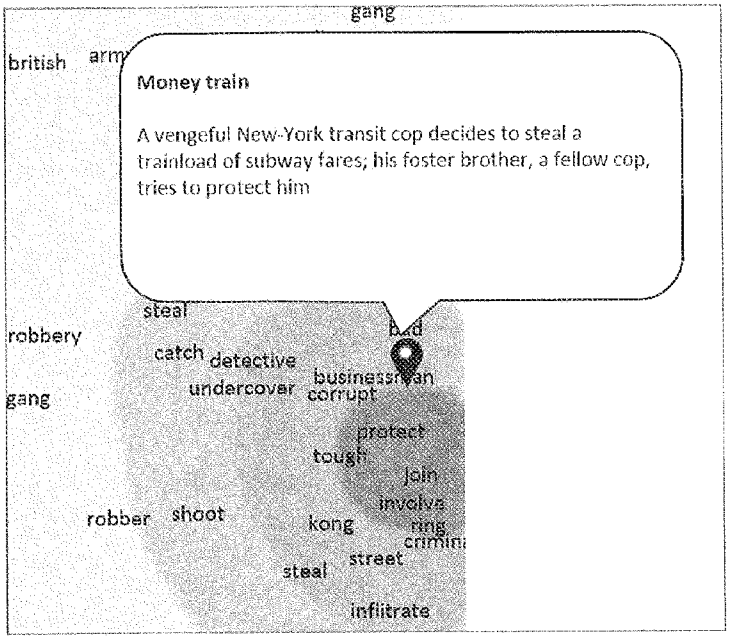

gang
british
Money train
A vengeful New-York transit cop decides to steal a trainload of subway fares; his foster brother, a fellow cop, tries to protect him
steal
robbery
catch
detective
undercover
businessman
corrupt
gang
protect
tough
join
involve
robber
shoot
kong
ring
street
steal
infiltrate

FIG. 15

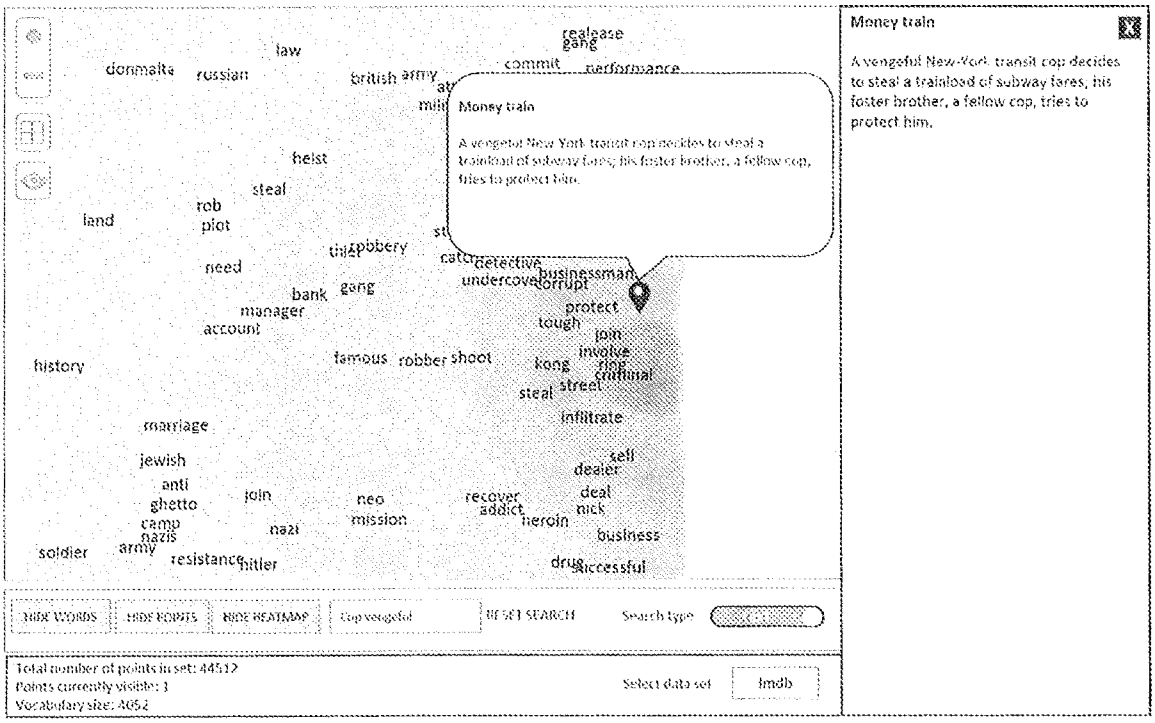

Money train
A vengeful New-York transit cop decides to steal a trainload of subway fares, his foster brother, a fellow cop, tries to protect him.
law
russian
commit
performance
heist
steal
rob
plot
land
need
bank
gang
manager
account
history
famous
robber
shoot
detective
undercover
businessman
corrupt
protect
tough
join
involve
kong
criminal
street
steal
infiltrate
marriage
jewish
anti
ghetto
join
neo
mission
recover
addict
deal
mick
heroin
business
camp
nazis
nazi
soldier
army
resistance
hitler
drug
successful
HIDE WORDS
HIDE POINTS
HIDE HEATMAP
RESET SEARCH
Search type
Total number of points in set: 44512
Points currently visible: 1
Vocabulary size: 4052
Select data set
Imdb

FIG. 16

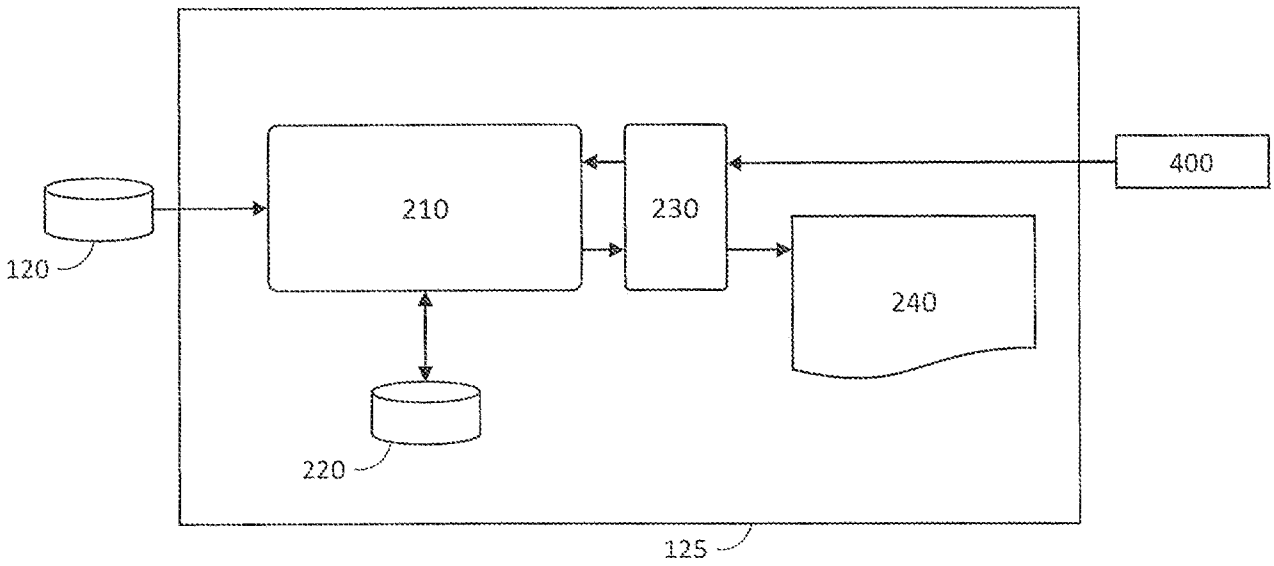

120
210
230
400
240
220
125

FIG. 17

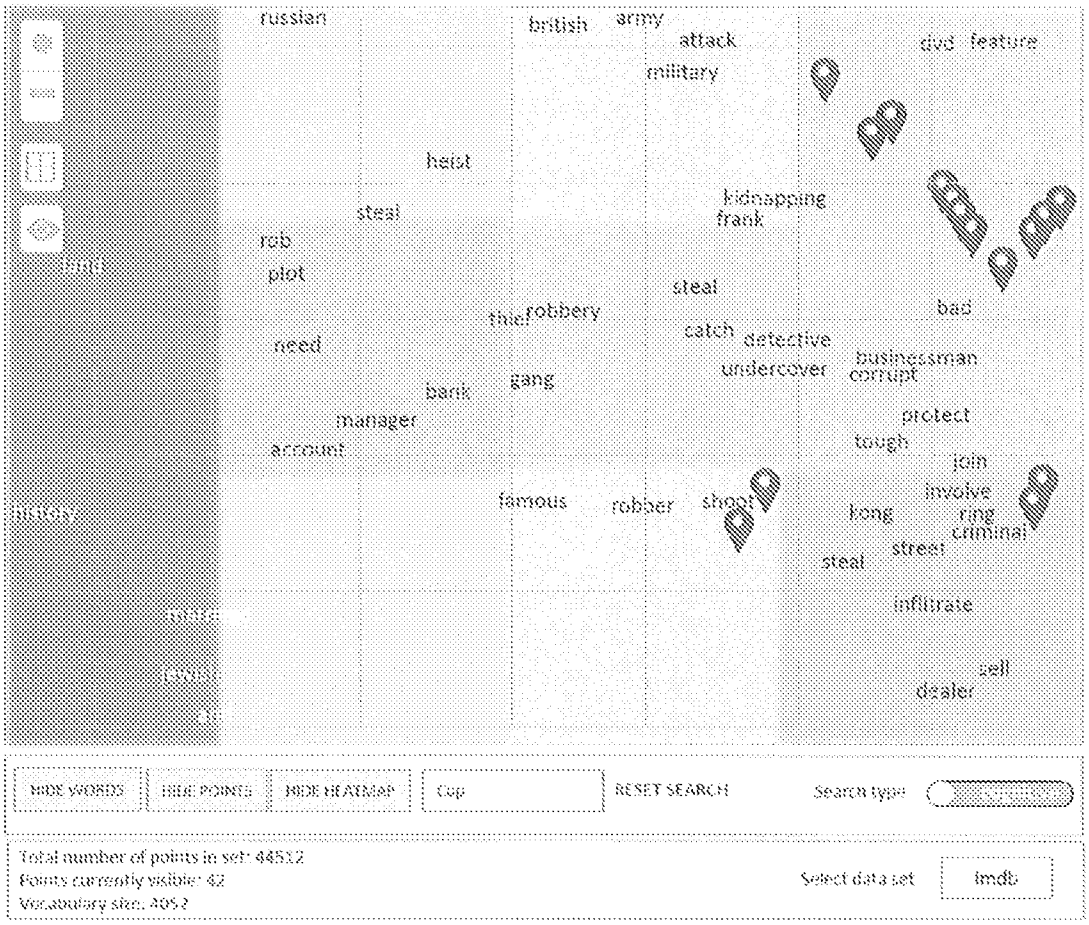

russian
british
army
attack
military
died
feature
heist
steal
kidnapping
frank
rob
plot
steal
robbery
bad
need
catch
detective
undercover
businessman
corrupt
bank
gang
protect
manager
account
tough
join
involve
famous
robber
shoot
long
ring
criminal
street
steal
infiltrate
sell
dealer
HIDE WORDS
HIDE POINTS
HIDE HEATMAP
cop
RESET SEARCH
Search type
Total number of points in set: 44512
Points currently visible: 42
Vocabulary size: 4052
Select data set
imdb

FIG. 18