(19) CONFÉDÉRATION SUISSE
INSTITUT FÉDÉRAL DE LA PROPRIÉTÉ INTELLECTUELLE

(11) CH 704 148 A2

(51) Int. Cl.: G06F 17/30 (2006.01)

**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: 02009/10

(22) Date de dépôt: 30.11.2010

(43) Demande publiée: 31.05.2012

(71) Requérant:
SALSAdev SA, succursale du Noirmont, Sous-la-Velle 14
2340 Le Noirmont (CH)

(72) Inventeur(s):
Nicolas Gamard, 1208 Genève (CH)
Stéphane Gamard, 1228 Plan-les-Ouates (CH)

(74) Mandataire:
SALSAdev SA Monsieur Nicolas GAMARD,
Chemin des Aulux 18
1228 Plan-les-Ouates (CH)

## (54) Procédé et système de recherche sémantique.

(57) L'invention concerne un procédé de recherche sémantique dans lequel un utilisateur désigne manuellement une partie de document comme étant pertinente pour une recherche, une recherche sémantique dans une ontologie particulière étant effectuée sur la base de la partie sélectionnée du document, les résultats de la recherche sémantique étant présentés à l'utilisateur.

## Description

### Domaine technique

**[0001]** L'invention concerne un procédé et un système de recherche sémantique. L'invention concerne en particulier un procédé et un système de recherche sémantique dans lequel un utilisateur désigne manuellement une partie d'un document comme étant pertinente pour la recherche sémantique.

### Technique antérieure

**[0002]** La recherche sémantique est une activité récente qui s'intéresse à la signification des mots et des expressions. Le terme de sémantique est souvent utilisé en opposition à celui de syntaxe. Il y a entre la sémantique et la syntaxe le même rapport qu'entre le fond et la forme.

**[0003]** EP 1 843 256 A1 décrit une méthode de classement d'entités en fonction de leurs contenus, leurs prépondérances étant influencées par un contexte défini. La recherche sémantique se fait sur la totalité du document. L'invention présente le désavantage que l'utilisateur ne peut pas améliorer la recherche en sélectionnant un sous-ensemble du document.

**[0004]** US 2010/0057702 décrit une méthode de génération de données à rechercher dans une pluralité de systèmes d'entreprises. Là également, l'invention présente le désavantage que l'utilisateur ne peut pas améliorer la recherche en sélectionnant un sous-ensemble du document.

**[0005]** US 7 099 870 décrit une méthode pour mettre à jour automatiquement et périodiquement une page web personnalisée en fonction de la détermination de la pertinence des informations contenues. L'invention présente le désavantage de ne pas avoir une interface utilisateur conviviale.

### Résumé de l'invention

**[0006]** Le but de la présente invention est de proposer un nouveau procédé et un nouveau système pour effectuer une recherche sémantique.

**[0007]** D'après la présente invention, ce but est atteint en particulier par les caractéristiques des revendications indépendantes. D'autres formes de réalisations avantageuses résultent en outre des revendications dépendantes et de la description.

**[0008]** Ce but est atteint en particulier par l'invention dans le fait que le procédé de recherche sémantique se compose de la désignation par un utilisateur d'un motif dans un document, de la recherche sémantique dans une ontologie sur la base du motif désigné et de la présentation à l'utilisateur du résultat de la recherche sémantique.

**[0009]** Dans le présent document, une ontologie désigne l'ensemble structuré des termes et concepts représentant le sens d'un élément d'informations, que ce soit par les métadonnées d'un espace de noms, ou les éléments d'un domaine de connaissances. L'ontologie constitue en soi un modèle de données représentatif d'un ensemble de concepts dans un domaine, ainsi que des relations entre ces concepts. Elle est employée pour raisonner à propos des objets du domaine concerné.

**[0010]** Dans le présent document, un marqueur désigne un mot-clef ou un terme associé ou assigné à de l'information, qui décrit ainsi l'objet et permet une classification des informations basée sur les mots-clefs.

**[0011]** Dans le présent document, un pointeur désigne un dispositif activé par l'utilisateur pour désigner une ou plusieurs parties de texte dans un document. Ce peut être un clavier, un trackball ou une souris associés à un environnement Windows, un capteur haptique associé à l'écran tactile d'un smartphone, un capteur de position d'œil associé à un projecteur rétinien au moyen duquel un document est directement projeté sur la rétine de l'utilisateur et les parties du document désignées sur la base de la direction d'observation de l'œil, un capteur vocal destiné à interpréter les commandes vocales émises par un utilisateur pour désigner une ou plusieurs parties d'un flux vocal.

**[0012]** Dans le présent document, un document désigne un ensemble de caractères ou de mots ou de motifs graphiques dans un quelconque format. Le document désigne également la conversion en texte ou en expressions d'un flux auditif. Le flux peut être saisi et converti en temps réel, par exemple dans le cas d'un microphone situé à proximité ou non de l'utilisateur. Un flux peut être saisi et converti en différé par exemple dans le cas de la bande son d'un film. Un flux peut être converti en texte dans le cas d'éléments bien définis de la communication verbale ou en marquages pour tenir compte d'éléments de la communication non verbale, comme des expressions («oh!») ou des hésitations («mhhh»). De même, des bruits ou bruitages peuvent être identifiés et donner lieu à des marquages particuliers («crissement de pneus», «galop de cheval»). De manière générale, un document peut contenir des marqueurs pour indexer certaines parties de son contenu. Par analogie, un document peut également concerner des informations adressées à d'autres sens humains que la vue et l'ouïe, comme par exemple le toucher, l'odorat, auxquels cas le marquage sera spécifique à ces sens.

**[0013]** Le but est atteint en particulier par l'invention dans le fait que le procédé de recherche sémantique se compose de la désignation par un utilisateur d'un motif dans un document, la désignation se faisant par exemple au moyen d'un pointeur, le document étant par exemple un document texte dans un quelconque format, par la recherche sémantique dans une ontologie par une application sur la base du contenu du motif désigné, l'ontologie étant établie sur la base d'une collection

de documents préalablement indexés et de la présentation à l'utilisateur du résultat de la recherche sémantique, le résultat de la recherche s'affichant par exemple dans une nouvelle fenêtre de son ordinateur, l'ensemble des éléments résultant de la recherche étant représentés par exemple sous la forme d'une énumération ou sous celle d'un plan à deux dimensions ou sous celle d'une sphère, permettant ainsi un affichage par ordre d'importance, ou sous la forme d'une information audible présentée à l'utilisateur. L'avantage de ce procédé réside dans le fait que la qualité des résultats présentés à l'utilisateur s'en trouve améliorée si la recherche se base prioritairement sur les motifs désignés par l'utilisateur et non pas sur la totalité du document initial.

### Brève description des dessins

**[0014]** Une réalisation de la présente invention est décrite dans ce qui suit à l'aide d'un exemple illustré par la figure unique annexée qui montre une session de travail à l'ordinateur représentant des fenêtres avec des documents et des résultats de recherche sémantique.

### Description des modes de réalisation

**[0015]** Dans un mode de réalisation préférentiel, le document 1 est un document de texte, une page web, un courriel, un système de fichiers, un serveur collaboratif (sharepoint), ou tout document composé d'un ensemble de caractères, de mots ou de pictogrammes. L'avantage de cette variante réside dans le fait que le système est indépendant du format des données et qu'il est possible d'effectuer une recherche parmi des documents de texte de formats différents.

**[0016]** Dans un mode de réalisation préférentiel, le document 1 est un flux de caractères qui peut être en temps réel ou non, comme des rss (Really Simple Syndication) ou des tweets. Ces flux d'informations sont généralement transportés de manière numérique par des réseaux de télécommunications comme Internet. Dans ce mode de réalisation, les flux peuvent être générés par des personnes, par des agents intelligents, à savoir des applications dotées d'une logique, par des objets dotés d'une interface de communication et capables d'émettre de tels flux. L'avantage de cette variante réside dans le fait que les sources d'informations peuvent être humaines ou non, reliées au monde réel ou non.

**[0017]** Dans un mode de réalisation préférentiel, la recherche sémantique peut se faire sur la base du motif 11 désigné par l'utilisateur ainsi et sur la base d'un champ marqué dans le document initial. Le champ concernera par exemple l'auteur du document, la date de sa création, de sa modification, de sa publication, de sa réception ou concernera le résumé du document. L'avantage de cette variante réside dans le fait qu'il est possible de préciser la recherche avec des éléments supplémentaires et, partant, d'en améliorer la qualité. Un autre avantage de cette variante réside dans le fait qu'il est possible de rejeter explicitement une partie du document que l'utilisateur sait non pertinente.

**[0018]** Dans un mode de réalisation préférentiel, l'utilisateur choisit de faire la recherche dans une ontologie particulière parmi plusieurs ontologies 2 qui lui sont présentées. Par exemple, une recherche sur le financement d'un programme de recherche donnera de meilleurs résultats si l'ontologie EUresearch 21 est utilisée au lieu d'une ontologie de politique 22 ou de finance 23. Dans cet exemple, l'ontologie EUresearch 21 regroupe une collection de projets dans le cadre de programmes nationaux européens de recherche. L'avantage de cette variante réside dans le fait que le résultat de la recherche est substantiellement amélioré par l'utilisation des connaissances implicites de l'utilisateur.

**[0019]** Dans un mode de réalisation préférentiel, le document 1 provient de la conversion d'un flux audio en texte. La conversion peut avoir lieu en temps réel, par exemple dans le cas d'un flux provenant d'un microphone associé à l'utilisateur et situé ou non dans son environnement immédiat. La conversion peut avoir lieu en différé, par exemple dans le cas du traitement de la bande son d'un film. Dans cette variante, il est possible d'effectuer une recherche sur la base de la bande sonore d'un film qui sera préalablement convertie en texte, au demeurant non structuré. L'avantage de cette variante réside dans le fait qu'il est possible d'étendre la recherche à des éléments de communication verbale ou non verbale.

**[0020]** Dans un mode de réalisation préférentiel, le motif 11 est un segment de texte. Le segment de texte 11 peut correspondre à un ensemble de mots ou de caractères continu ou discontinu. L'avantage de ce mode de réalisation réside dans le fait que la désignation d'un segment continu de mots ou de caractères peut se faire rapidement et que la désignation d'un segment discontinu sous la forme d'un ensemble de mots ou de caractères peut gagner en précision en excluant explicitement des mots qui pourraient perturber la recherche. La désignation d'un ensemble discontinu de mots peut se faire, sous le système d'exploitation Windows, en maintenant pressée la touche Ctrl lors de la désignation des différentes parties de l'ensemble au moyen de la souris.

**[0021]** Dans un mode de réalisation préférentiel, le motif 11 est désigné par l'utilisateur au moyen d'un dispositif de pointage comme une souris, un trackball, au moyen du déplacement de doigts sur un écran tactile, au moyen de la détermination de la direction du regard de l'utilisateur sur le document. L'avantage de ce mode de réalisation réside dans le fait que l'utilisation intuitive par l'utilisateur des interfaces usuelles qui équipent son PC ou son smartphone améliore la simplicité et l'efficacité de la préparation de la recherche.

**[0022]** Un utilisateur intéressé, par exemple un membre d'une promotion économique ou d'un institut d'aide à la recherche, peut effectuer une recherche sur la base du contenu d'un courriel 1 reçu d'une haute école intéressée à créer un nouveau projet et à trouver un moyen de financement pour ce projet. Le courriel 1 décrit plus ou moins en détail le nouveau projet, les buts et les attentes ainsi que la durée et le budget envisagés. Les parties pertinentes 11 du courriel 1 sont entourées

d'informations qui ne sont pas pertinentes pour la recherche comme des formules de politesse 12 ou des informations d'exclusion (disclaimer) non représentées sur la figure. A l'aide d'un dispositif de pointage, l'utilisateur marque la partie 11 du courriel 1 qui correspond à la description du nouveau projet et qu'il juge pertinente. En s'aidant de la touche ctrl du clavier, l'utilisateur peut au besoin désigner un motif 11 qui n'est pas d'un seul tenant mais composé de plusieurs parties désignées séquentiellement. Une fois le motif 11 désigné dans le document 1, celui-ci est glissé à l'aide de la souris sur une nouvelle fenêtre 2, en particulier sur une icône 21 représentant l'ontologie EUresearch 21 correspondant à tous les programmes de recherche de l'Union Européenne préalablement indexés par le moteur de recherche sémantique selon des règles définies. La recherche s'effectue alors sur la base du motif 11 désigné. L'affichage du résultat de la recherche se fait dans une nouvelle fenêtre 3, par exemple sous la forme de la liste des programmes de recherche répondant aux critères, éventuellement énumérés selon un ordre de pertinence, la pertinence du résultat étant déterminée dans une métrique multi-vecteurs par la distance entre la trace du motif 11 désigné et chacun des documents de référence préalablement indexés dans l'ontologie.

**[0023]** Dans une variante de réalisation, la recherche sémantique se fait sur la base d'un flux audio (document) dont la source est un microphone associé à l'utilisateur. Le microphone est personnel, mobile et monté directement sur l'équipement de l'utilisateur, de manière visible ou non. Le microphone capte l'environnement sonore de l'utilisateur d'une manière continue, convertit les conversations captées en texte et les expressions non verbales en marqueurs. Au moyen de commandes vocales, l'utilisateur désigne les parties du flux qu'il considère comme étant pertinentes. Celles-ci sont analysées sémantiquement, par exemple en regard de flux antérieurs accumulés sur la base du microphone et constituant une ontologie. Le microphone peut être à plusieurs canaux pour capter, par exemple, une information en stéréo ou spatiale. Les informations issues de la recherche sémantique peuvent être restituées à l'utilisateur immédiatement ou ultérieurement, par exemple sur demande. Dans le cas où elles sont restituées immédiatement, elles peuvent l'être sous une forme auditive par l'intermédiaire d'une oreillette ou sous une forme optique par l'intermédiaire d'un écran ou d'un projecteur rétinien. L'avantage de cette variante de réalisation réside dans le fait que l'analyse sémantique tient compte de tout l'environnement sonore de l'utilisateur et ne se limite pas seulement à la partie que le conscient de l'utilisateur a retenu après filtrage des éléments qu'il ne considérait pas comme pertinents.

**[0024]** Dans une variante de réalisation, la recherche sémantique se fait sur la base d'un indexage sémantique de nature mathématique et qui consiste en seulement 1% de la taille du document original. Une fois les documents indexés, ils sont disponibles notamment pour une recherche par mots-clefs, pour une découverte, pour une exploration de données.

**[0025]** Dans une variante de réalisation, la recherche sémantique se fait sur la base d'un document ou parties d'un document.

**[0026]** Dans une variante de réalisation, la recherche sémantique se fait sous la forme d'un marquage automatique et d'une classification. Dans le cas d'un marquage automatique, les marqueurs sont extraits automatiquement des documents. Dans le cas d'une classification, les documents sont automatiquement caractérisés et assignés à un membre d'une collection de catégories préalablement définies par l'utilisateur.

**[0027]** Dans une variante de réalisation, le résultat de la recherche sémantique apparaît sous la forme d'une carte de connaissances. Cette carte peut alors être utilisée pour explorer, sonder de manière intelligente une masse de documents non structurés. Application industrielle.

**[0028]** L'application industrielle est avérée dans la mesure où la recherche sémantique va jouer un rôle toujours plus grand dans le Web sémantique. Il y a un besoin substantiel de recherche sémantique dans de nombreux domaines commerciaux, notamment dans le domaine de la gestion de contrats, le domaine légal et le domaine financier.

## Revendications

1. Procédé de recherche sémantique, se composant des étapes suivantes:
   - désignation par un utilisateur d'un motif dans un document,
   - recherche sémantique dans une ontologie par une application sur la base du motif désigné,
   - présentation à l'utilisateur du résultat de la recherche sémantique.
2. Procédé selon la revendication 1 dans lequel le motif est un segment de texte.
3. Procédé selon la revendication 1 dans lequel l'utilisateur désigne le motif au moyen d'un pointeur.
4. Système de recherche sémantique, se composant des éléments suivants:
   - une interface utilisateur au moyen de laquelle un utilisateur désigne un motif dans un document,
   - un module au moyen duquel une recherche sémantique dans une ontologie est effectuée sur la base du motif désigné,
   - une interface utilisateur au moyen de laquelle le résultat de la recherche sémantique est présenté.
5. Dispositif (Serveur) de recherche sémantique comprenant
   - une interface pour la réception d'un motif désigné par un utilisateur dans un document,
   - un module au moyen duquel une recherche sémantique dans une ontologie est effectuée sur la base du motif désigné,

- une interface pour la présentation du résultat de la recherche sémantique.

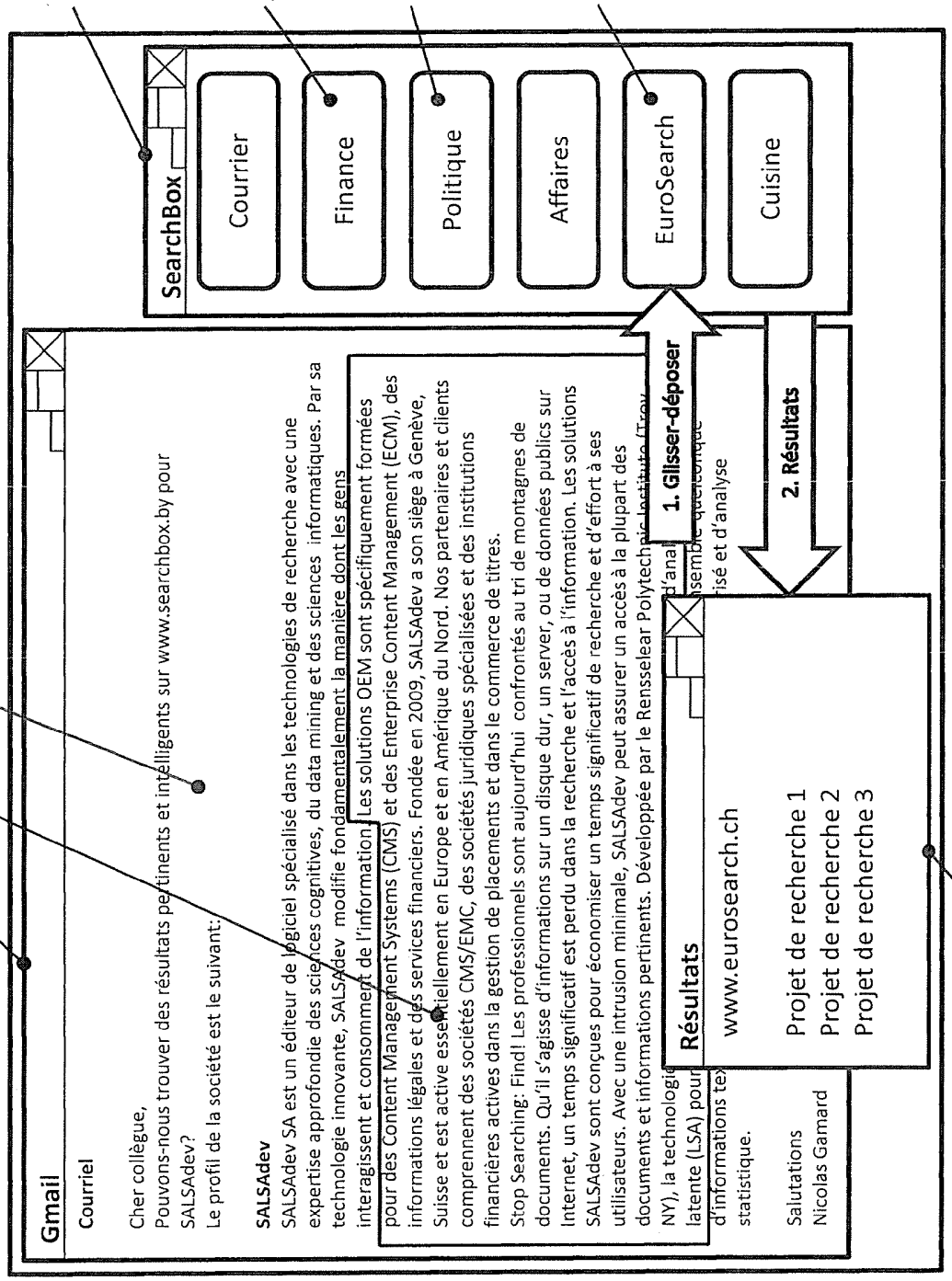
1
11
12
2
23
22
21
3
Gmail
Courriel
Cher collègue,
Pouvons-nous trouver des résultats pertinents et intelligents sur www.searchbox.by pour SALSAdev?
Le profil de la société est le suivant:
SALSAdev
SALSAdev SA est un éditeur de logiciel spécialisé dans les technologies de recherche avec une expertise approfondie des sciences cognitives, du data mining et des sciences informatiques. Par sa technologie innovante, SALSAdev modifie fondamentalement la manière dont les gens interagissent et consomment de l'information. Les solutions OEM sont spécifiquement formées pour des Content Management Systems (CMS) et des Enterprise Content Management (ECM), des informations légales et des services financiers. Fondée en 2009, SALSAdev a son siège à Genève, Suisse et est active essentiellement en Europe et en Amérique du Nord. Nos partenaires et clients comprennent des sociétés CMS/EMC, des sociétés juridiques spécialisées et des institutions financières actives dans la gestion de placements et dans le commerce de titres.
Stop Searching: Find! Les professionnels sont aujourd'hui confrontés au tri de montagnes de documents. Qu'il s'agisse d'informations sur un disque dur, un server, ou de données publics sur Internet, un temps significatif est perdu dans la recherche et l'accès à l'information. Les solutions SALSAdev sont conçues pour économiser un temps significatif de recherche et d'effort à ses utilisateurs. Avec une intrusion minimale, SALSAdev peut assurer un accès à la plupart des documents et informations pertinents. Développée par le Rensselear Polytechnic Institute (Troy NY), la technologie latente (LSA) pour
d'informations tex
statistique.
Salutations
Nicolas Gamard
SearchBox
Courrier
Finance
Politique
Affaires
EuroSearch
Cuisine
1. Glisser-déposer
2. Résultats
Résultats
www.eurosearch.ch
Projet de recherche 1
Projet de recherche 2
Projet de recherche 3

Fig.