(12) DEMANDE INTERNATIONALE PUBLIÉE EN VERTU DU TRAITÉ DE COOPÉRATION EN MATIÈRE DE BREVETS (PCT)

(19) Organisation Mondiale de la Propriété Intellectuelle
Bureau international

(43) Date de la publication internationale
28 juillet 2005 (28.07.2005)

PCT

(10) Numéro de publication internationale
WO 2005/069166 A1

(51) Classification internationale des brevets[7] : G06F 17/30

(21) Numéro de la demande internationale :
PCT/FR2005/000023

(22) Date de dépôt international : 6 janvier 2005 (06.01.2005)

(25) Langue de dépôt : français

(26) Langue de publication : français

(30) Données relatives à la priorité :
0400060 6 janvier 2004 (06.01.2004) FR

(71) Déposant *(pour tous les États désignés sauf US)* : **CORBIERE, Jérôme** [FR/FR]; 37, Rue de Nalpes, F-75008 Paris (FR).

(71) Déposant et

(72) Inventeur : **DURIF, Sylvain** [FR/FR]; 21, Boulevard de l'Hippodrome, F-01000 Bourg En Bresse (FR).

(74) Mandataires : **PEAUCELLE, Chantal** etc.; Cabinet Armengaud Aine, 3, Avenue Bugeaud, F-75116 Paris (FR).

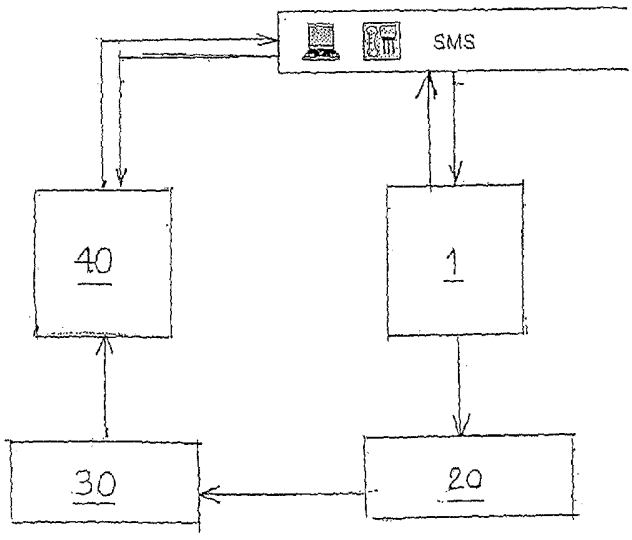
(81) États désignés *(sauf indication contraire, pour tout titre de protection nationale disponible)* : AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) États désignés *(sauf indication contraire, pour tout titre de protection régionale disponible)* : ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), eurasien (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), européen (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, MC, NL, PL, PT, RO,

[Suite sur la page suivante]

(54) Title: AUTOMATIC SYSTEM FOR RETRIEVING AND PROCESSING INFORMATION CARRIED BY SHORT MESSAGES

(54) Titre : SYSTEME AUTOMATIQUE DE TRAITEMENT DES INFORMATIONS PORTEES PAR DES TEXTES COURTS



(57) **Abstract:** The invention relates to an automatic system for retrieving and processing information carried by short messages on all types of communications carriers and for carrying out a target dissemination of search results concerning relations between said information or comparisons therebetween on all types of communications carriers. The inventive system comprises a module (1) for receiving and pre-decoding initial short messages, a module (20) for pre-processing said messages, a module (30) for retrieving information carried by said messages and a module (40) for using said information and for carrying out the target dissemination of search results concerning relations between said information or the comparisons therebetween.

[Suite sur la page suivante]

WO 2005/069166 A1

SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Publiée :

— *avec rapport de recherche internationale*

— *avant l'expiration du délai prévu pour la modification des revendications, sera republiée si des modifications sont reçues*

En ce qui concerne les codes à deux lettres et autres abréviations, se référer aux "Notes explicatives relatives aux codes et abréviations" figurant au début de chaque numéro ordinaire de la Gazette du PCT.

(57) Abrégé : Système automatique d'extraction et de traitement des informations portées par des textes courts transmis sur tous supports de communications et de diffusion ciblée sur tous supports de communications des résultats de recherche de relations entre ces informations ou de comparaisons établies entre elles. Ce système comprend : - un module de réception et de prédécodage (1) des textes cours initiaux, - un module de prétraitement (20) desdits textes, - un module d'extraction (30) des informations portées par lesdits textes, - un module d'exploitation desdites informations et de diffusion ciblée (40) des résultats de recherche de relations entre lesdites informations ou de comparaisons établies entre elles. (Dessin spécial)

« Système automatique de traitement des informations portées par des textes courts »

L'invention se rapporte, de façon générale, au traitement de textes courts constituant un corpus très souvent dégradé afin d'en extraire les informations qu'ils portent et, de façon plus particulière, à la recherche de relations ou à l'établissement de comparaisons entre les informations portées par différents textes courts et à la diffusion ciblée des résultats obtenus, la transmission desdits textes courts utilisant tous supports de communications, notamment les réseaux de téléphonie fixe ou mobile.

On entendra par "textes courts" des textes essentiellement composés de quelques phrases à structure simple, éventuellement de quelques paragraphes et auxquels peuvent être associés des données audio et/ou vidéo et/ou photographiques. Des petites annonces, des télégrammes, des informations fil rouge, des messages vocaux ou écrits peuvent, par exemple, constituer des textes courts.

Il existe déjà des procédés qui permettent de transformer des textes bruts en textes normalisés, adaptés à des services particuliers comme, par exemple, celui décrit dans le brevet US N° 6321372 de Hervé Poirier et coll. intitulé "Executable for requesting a linguistic service".

Le traitement de données à base d'automates informatiques d'état fini a fait l'objet de plusieurs travaux comme, par exemple, ceux présentés dans le brevet US N° 5564058 de Ronald M. Kaplan et coll. intitulé "Stored string data with encoded data units from subranges of values that indicate search information".

L'art antérieur le plus proche est constitué, en particulier, par un article du MIT Laboratory for Computer Science, Cambridge USA (Helen MENG et coll.) portant sur un système fondé sur un dialogue essentiellement vocal qui fournit un accès à une base de données de petites annonces dans le domaine automobile. L'art antérieur comporte également le brevet international WO 01/63451 déposé par CLASSIFIED ADVERTISER COM, INC. traitant d'un système de commande et de placement d'annonces à distance et le brevet européen EP 1185062 déposé par NETCALL PLC décrivant l'établissement d'une connexion téléphonique entre un client et un annonceur.

Les inconvénients majeurs de cet art antérieur résident dans le fait qu'il se limite essentiellement,

soit au traitement de petites annonces, alors que les textes dits courts peuvent trouver un usage dans d'autres domaines que celui des petites annonces,

soit à un traitement sur un seul support de communications, alors que les textes courts sont susceptibles d'être acheminés sur tous les supports de communications.

De plus, cet art antérieur ne nous apprend sensiblement rien sur les textes courts considérés comme constituant un corpus dégradé, corpus dégradé par l'emploi de diverses abréviations ou par la présence d'erreurs typographiques ou orthographiques, lequel nécessitera des traitements spécifiques pour en extraire de façon automatique l'information qu'il porte.

Les systèmes proposés jusqu'ici n'envisagent pas un traitement automatique global portant sur les contenus de textes courts transmis sur tous supports de communications et concernant tous domaines.

Ils n'envisagent pas non plus l'établissement de relations croisées entre des informations portées par des textes relatifs à des domaines d'intérêt différents.

Le premier aspect de l'invention qui tend essentiellement à combler les manques constatés dans l'art antérieur concerne un système automatique d'extraction et de traitement des informations portées par des textes courts, accompagnés ou non de données audio et/ou vidéo et/ou photographiques, transmis sur tous supports de communications et de diffusion ciblée sur tous supports de communications des résultats de recherche de relations entre ces informations ou de comparaisons établies entre elles, caractérisé en ce que ledit système est rendu automatique par la mise en oeuvre d'un serveur équipé d'un microprocesseur qui commande et organise la combinaison des modules suivants, comprenant :

A/ au moins un premier module, dit de réception et de précodage, comprenant :

- des moyens de récepteur destinés à recevoir, de façon continue ou non, une pluralité de textes courts transmis sur un quelconque support de communications, chacun d'entre eux étant associé à un premier code identifiant l'utilisateur émetteur d'un texte court, codé en binaire par des premiers moyens de codeur primaire,

- des seconds moyens de codeur binaire destinés à identifier sous forme codée la nature du support de communications acheminant ledit texte court,

- des moyens de numérisation connectés aux moyens de récepteur et destinés à numériser les contenus desdits textes courts non préalablement numérisés par le mode de transmission,

- des premiers moyens d'horloge associés à des troisièmes moyens de codeur binaire destinés à coder des données d'horodatage concernant la réception desdits textes courts,

- des moyens d'identification et d'authentification de l'émetteur dudit texte court par vérification du premier code qui y est associé,

- des moyens de reconnaissance de la langue dans laquelle ledit texte court est écrit, associés à des quatrièmes moyens de codeur binaire pour identifier sous forme codée la langue reconnue,

- des moyens de reconnaissance du domaine concerné par le contenu dudit texte court, associés à des cinquièmes moyens de codeur binaire pour identifier, sous forme codée, le domaine reconnu,

- une pluralité de bases de données connectée respectivement aux moyens d'identification, aux moyens de reconnaissance de la langue et du domaine,

- des premiers moyens de registre recevant les différents codes, de façon à constituer, sous forme d'un mot binaire, une cellule établie pour être associée en permanence à un et un seul texte court reçu,

- des moyens de concaténation de chaînes binaires destinés à ajouter la cellule à la chaîne de sortie des moyens de reconnaissance de domaine, pour obtenir une chaîne unique regroupant texte court et cellule,

- des moyens de transmission, en retour vers l'utilisateur d'origine, du texte court reçu par le serveur converti en langage clair par des moyens de conversion et accompagné de la cellule associée, cette dernière étant décodée dans des moyens de décodeur ,

- des moyens de validation dépendant du support de communications utilisé et destinés à délivrer un signal de validation généré par ledit utilisateur émetteur de texte court en réponse à son approbation du texte et de la cellule qui lui ont été retransmis, toute modification de la cellule devenant impossible après ladite validation,

- des moyens de circuit logique recevant en première entrée la chaîne unique produite par les moyens de concaténation et ne pouvant délivrer, en sortie, cette même chaîne que si le signal de validation est appliqué en seconde entrée,

- des premiers moyens de mémoire destinés à stocker ladite chaîne unique produite par les moyens de concaténation regroupant texte court et cellule, l'accès aux dits premiers moyens de mémoire étant commandés par lesdits moyens de circuit logique ;

B/ au moins un second module , dit de prétraitement des textes, comprenant:

- des moyens de segmentation destinés à séparer les phrases formant ledit texte court et associés à des premiers moyens de compteur réservés à la numérotation des différentes phrases reconnues, et à la numérotation de chaque caractère utilisé pour former les phrases dudit texte court,

- des moyens de découpage destinés à séparer les mots d'une phrase reconnue, et associés à des seconds moyens de compteur réservés à la numérotation attribuée à chacun des caractères de début et de fin de chaque mot par rapport à l'ensemble des mots identifiés dans une phrase reconnue,

- des moyens d'analyse morphologique et morpho-syntaxique de tous les mots identifiés d'une phrase, destinés à transformer chaque mot identifié en sa forme normalisée (lemme) et à lui associer des attributs liés à sa catégorie grammaticale exacte en fonction de son contexte proche,

- des premiers moyens de file d'attente destinés à stocker les différentes phrases identifiées et numérotées en sortie desdits moyens de segmentation ,

- des seconds moyens de file d'attente destinés à stocker les différents mots identifiés dont les caractères ont été numérotés en sortie desdits moyens de découpage ,

- des troisièmes moyens de file d'attente destinés à stocker les mots identifiés normalisés et leurs attributs associés en sortie desdits moyens d'analyse morphologique et morpho-syntaxique ,

- des seconds moyens de mémoire destinés à stocker en sortie des troisièmes moyens de file d'attente une pluralité de fiches d'analyse constituées, chacune, des mots identifiés normalisés et de leurs attributs associés correspondant à une phrase donnée, chaque fiche d'analyse étant accompagnée de la cellule

correspondant audit texte court d'origine, cette dernière étant invariante au cours des traitements cités,

- des seconds moyens d'horloge destinés à délivrer, selon des séquences pré-programmées, des signaux de commande d'accès, respectivement, aux moyens de traitement et de mémoire ;

C/ au moins un troisième module, dit d'extraction des informations, comprenant:

- une pluralité de dispositifs d'extraction des informations, chacun étant spécifique d'un domaine, ces dispositifs étant destinés à transformer les attributs associés à chaque mot de la fiche d'analyse en de nouveaux attributs identifiant l'information portée par chaque mot, convertissant ainsi chaque fiche d'analyse en une fiche préliminaire,

- des troisièmes moyens de mémoire destinés à stocker lesdites fiches préliminaires et leur cellule associée qui reste invariante en traversant un dispositif d'extraction quelconque,

- au moins un système de bases de données comportant, pour chaque domaine spécifique traité par un dispositif d'extraction donné, une base élémentaire de données, vide d'enregistrement, constituée uniquement d'une pluralité de noms de champ qui définit une fiche type, le système de bases de données comportant également son système de gestion et son dispositif de mémorisation,

- des moyens de chargement des données contenues dans une fiche préliminaire, dans ladite base élémentaire de données, les nouveaux attributs de ladite fiche préliminaire correspondant aux noms de champ de la fiche type associée, le résultat étant l'établissement d'une fiche caractéristique organisée de façon prédéterminée et destinée à identifier l'objet et ses paramètres associés de l'information portée par chaque phrase d'un texte court reçu,

- des quatrièmes moyens de file d'attente destinés à stocker de façon temporaire lesdites fiches caractéristiques issues dudit système de bases de données,

- des quatrièmes moyens de mémoire connectés aux quatrièmes moyens de file d'attente et destinés à stocker de façon permanente lesdites fiches caractéristiques et les cellules associées,

- des troisièmes moyens d'horloge destinés à délivrer, selon une séquence pré-programmée, des signaux de commande des accès, respectivement, à la pluralité

de dispositifs d'extraction, audit système de bases de données et aux dits quatrièmes moyens de mémoire ;

D/ au moins un quatrième module, dit d'exploitation et de diffusion ciblée, comprenant :

- des moyens de sélection, connectés aux quatrièmes moyens de mémoire, destinés à sélectionner une ou plusieurs fiche(s) caractéristique(s) et à les transférer dans une table à consulter

- au moins un réseau à logique programmable destiné à comparer sur des critères préprogrammés une fiche caractéristique courante provenant, soit des quatrièmes moyens de file d'attente, soit des quatrièmes moyens de mémoire avec une ou plusieurs autre(s) fiche(s) caractéristique(s) présente(s) dans ladite table à consulter et à délivrer en sortie le contenu de la (ou des) fiche(s) caractéristique(s) identifiée(s) comme répondant positivement à ladite comparaison,

- des seconds moyens de registre destinés, chacun, à stocker temporairement le contenu d'une fiche caractéristique identifiée comme répondant positivement à ladite comparaison et délivrée par ledit réseau à logique programmable,

- des moyens de balayage destinés à analyser successivement les sorties respectives desdits moyens de registre et à transférer les contenus de ces derniers,

- des cinquièmes moyens de file d'attente destinés à stocker temporairement les fiches caractéristiques identifiées positivement provenant desdits moyens de balayage ,

- des troisièmes moyens de compteur connectés aux dits seconds moyens de registre et destinés à compter le nombre de fiches caractéristiques identifiées comme répondant positivement à la comparaison exécutée par ledit réseau à logique programmable ,

- des cinquièmes moyens de mémoire destinés à stocker à des fins de sauvegarde, d'archivage ou d'études ultérieures, les fiches caractéristiques identifiées délivrées par les cinquièmes moyens de file d'attente ainsi que les cellules associées,

- des moyens de diffusion destinés à une diffusion ciblée sur le support de communication approprié des fiches caractéristiques identifiées positivement provenant, dans un ordre préprogrammé, des cinquièmes moyens de file d'attente ou

des cinquièmes moyens de mémoire, ces moyens de diffusion étant équipés de moyens de décodage et de conversion des données numériques, ces derniers moyens s'appliquant au contenu des fiches caractéristiques identifiées positivement et aux cellules associées,

- des quatrièmes moyens d'horloge destinés à horodater la diffusion ciblée exécutée par lesdits moyens de diffusion et à délivrer des signaux de commande d'accès, respectivement, audit réseau à logique programmable, aux dits moyens de balayage, aux dits moyens de diffusion et aux dits cinquièmes moyens de mémoire .

Dans un second aspect, l'invention décrit un procédé de traitement automatique de textes courts mettant en oeuvre le système décrit ci-dessus et caractérisé par les étapes suivantes consistant à :

a) entrer en communication avec le module de réception d'un serveur auquel sont transmis un texte court, accompagné ou non de données audio et/ou vidéo et/ou photographiques et un code utilisateur, ledit module numérisant, si nécessaire, les données reçues, vérifiant l'identification de l'utilisateur, et horodatant la réception dudit texte court ;

b) déterminer la langue et le domaine concernés par ledit texte court reçu ;

c) faire valider par l'utilisateur le contenu du texte reçu, la langue et le domaine reconnus ;

d) séparer les phrases du texte et les mots de chacune des phrases ;

e) pratiquer sur chaque mot une analyse morphologique et morpho-syntaxique dont le résultat est de produire un mot normalisé auquel est associé un attribut ;

f) extraire à partir des moyens d'extraction spécifiques du domaine reconnu à l'étape b l'information portée par chaque mot associé à son attribut pour convertir ce dernier en un ou plusieurs nouveau(x) attribut(s) parmi lesquels sont reconnus l'objet dudit texte court et les paramètres liés à cet objet ;

g) mémoriser chaque mot d'une phrase dudit texte court reçu avec ses nouveaux attributs dans une fiche préliminaire ;

h) obtenir une fiche caractéristique, en créant un enregistrement dans une fiche type constituée d'un champ objet et d'une pluralité de champs de paramètres à partir des données contenues dans une fiche préliminaire ;

i) mémoriser l'ensemble des fiches caractéristiques ;

j) comparer une ou plusieurs fiche(s) caractéristique(s) à une ou plusieurs fiche(s) caractéristique(s) sélectionnée(s) ;

k) sortir les résultats de chaque comparaison effectuée et les mémoriser ;

l) diffuser les résultats mémorisés et convertis vers l'utilisateur émetteur dudit texte court reçu ou vers tous autres destinataires autorisés.

Dans un troisième aspect, l'invention décrit plus particulièrement un procédé de traitement de textes d'annonces dont le contenu concerne des demandes ou des offres de produits et/ou de services mettant en oeuvre le système et le procédé décrits précédemment., caractérisé en ce que les comparaisons réalisées entre fiches caractéristiques portent uniquement sur les champs de paramètres de fiches caractéristiques dont les objets sont complémentaires.

Les avantages et les caractéristiques de l'invention apparaîtront plus clairement à la lecture de la description suivante d'un mode préféré de réalisation en relation avec les dessins annexés dans lesquels :

La figure 1 représente un schéma fonctionnel du premier module dit de réception et de précodage selon la présente invention,

La figure 2 représente un schéma fonctionnel du second module dit de prétraitement des textes selon la présente invention,

La figure 3 représente un schéma fonctionnel du troisième module dit d'extraction des informations selon la présente invention, et

La figure 4 représente un schéma fonctionnel du quatrième module dit d'exploitation et de diffusion ciblée selon la présente invention.

DESCRIPTION DU MODE PREFERE DE REALISATION

Dans la partie haute de la figure 1 représentant le module de réception et de précodage 1 selon la présente invention, un certain nombre non limitatif de supports de communications sont représentés de façon symbolique. Ces supports, notamment les réseaux de téléphonie fixe ou mobile, sont à la disposition des utilisateurs du système selon l'invention pour qu'ils puissent émettre, vers le récepteur 2 dudit module 1, un ou plusieurs textes courts après appel d'un numéro de serveur dédié suivi d'un code d'identification, ces textes courts étant, ou non, accompagnés de données audio et/ou vidéo et/ou de photographies. Ce code sera vérifié par des

moyens d'identification et d'authentification 4, ces derniers étant reliés à une première base de la pluralité de bases de données 11 relative aux utilisateurs du système.

Un premier codeur binaire 4' fournit le code d'identification (code 1) sous forme binaire et le transmet à des premiers moyens de registre 10.

Un second codeur binaire 2' reçoit du récepteur 2 l'indication concernant le support de communication sur lequel un texte court a été transmis et émet en conséquence le code binaire (code 2) approprié qu'il transmet aux dits premiers moyens de registre 10.

L'arrivée du texte court à l'entrée du récepteur est horodatée par les premiers moyens d'horloge 3 et les données d'horodatage sont transformées sous forme de code binaire (code 3) par un troisième codeur 3' et transmis aux premiers moyens de registre 10.

Les données de textes reçues par le récepteur 2 sont numérisées, si nécessaire, par des moyens classiques comme un convertisseur analogique-numérique (A/N), et sont transmises à des moyens de reconnaissance de la langue 5 utilisée pour écrire ledit texte reçu. En cas de textes courts transmis sur un réseau de téléphonie fixe ou mobile, un serveur vocal classique peut assurer la numérisation requise.

Ces moyens de reconnaissance 5 fonctionnent en utilisant le procédé connu per se des trigrammes et des mots courts (procédé TRISHORT).

Le modèle d'identification a été mis au point en analysant statistiquement des documents dans l'ensemble des langues reconnues.

Pour chaque langue, on détermine la valeur statistique de l'apparition de trois lettres consécutives.

Pour chaque langue, on détermine les fréquences d'apparition des mots communs courts (cinq lettres ou moins) comme "le", "et", "avec" ou "sinon". Par exemple, fondé sur le procédé précédent, le moteur linguistique XeLDA® multilingue mis au point par la société TEMIS (PARIS) permet de reconnaître 39 langues.

Pour ce faire, les moyens 5 doivent disposer d'un certain nombre de données qui sont stockées dans une seconde base de la pluralité de bases de données 11. Une fois identifiée, la langue est codée en binaire dans un quatrième codeur 5' et le code obtenu (code 4) est transféré dans lesdits premiers moyens de registre 10.

Une autre information importante est déterminée dans ce module 1, elle touche au domaine concerné par le contenu du texte comme, par exemple, le domaine bancaire, le domaine de l'immobilier, le domaine automobile, etc... Les moyens de reconnaissance du domaine 6 fonctionnement en faisant appel au procédé de catégorisation avec apprentissage.

Pour chaque domaine que l'on souhaite reconnaître, il est nécessaire, au préalable, de disposer d'un ensemble de textes le décrivant. Cet ensemble ou "corpus" est analysé statistiquement pour créer un profil de domaine composé d'une série de mots du domaine pondérés par des coefficients numériques.

Ainsi, un ensemble de domaines peut être associé à un autre ensemble de profils de domaine.

Enfin, pour déterminer si un nouveau document appartient à l'un ou l'autre des domaines, on crée, par une analyse linguistique de surface, un profil du document qui est comparé aux profils de domaine. Le domaine du document est celui dont le profil est le plus proche du profil du document.

Pour ce faire, il existe des outils informatiques tels que des descripteurs sémantiques déterminant dans un document donné les fréquences d'apparition de noms, de verbes, de groupes nominaux... Un algorithme permet ensuite de fusionner les différents descripteurs sémantiques attribués à une même catégorie pour créer un modèle de catégorisation.

Fonctionnant sur ce principe, le serveur de catégorisation mis au point par la Société TEMIS sous le nom d'Insight Discoverer™ Categorizer est tout à fait approprié pour constituer les moyens de reconnaissance de domaine 6 selon la présente invention, qui sont, de plus, connectés à une troisième base de la pluralité de bases de données 11 stockant, notamment, les différents profils susdits utiles à l'identification de domaine.

Le résultat de cette identification de domaine est codé en binaire par un cinquième codeur 6' et le code obtenu (code 5) est également transmis aux moyens de registre 10.

En sortie des moyens de reconnaissance de domaine 6, le texte court reçu est récupéré et on lui associe le mot binaire construit à partir des codes 1, 2, 3, 4

et 5 stockés dans lesdits moyens de registre 10 et définissant une cellule Ct spécifique d'un texte court reçu.

Ce mot binaire de la cellule est concaténé à la chaîne correspondant au texte court lui-même dans des moyens classiques de concaténation 12.

En sortie des moyens de concaténation, la chaîne unique (texte + cellule) est transmise à une première entrée d'un circuit logique 8 de fonction ET et également à des moyens de transmission 7 équipés de décodeurs 7' et de convertisseurs 7" destinés à réémettre en langage clair sur le support de communications approprié vers l'utilisateur à l'origine du texte court reçu, le contenu de ce dernier et les informations recueillies (horodatage, langue, domaine). L'utilisateur peut alors valider l'ensemble de ces informations à l'aide de moyens classiques de validation dépendant du support de communications sélectionné.

En cas de non validation, l'utilisateur peut réinitialiser son appel du serveur et reprendre à son point de départ le processus décrit ci-dessus. En cas de validation, le signal de validation V est transmis à une seconde entrée du circuit logique 8. Dès l'application du signal V, le circuit logique 8 transmet aux premiers moyens de mémoire 9 la chaîne correspondant à l'ensemble texte et cellule à des fins, soit de stockage, soit d'attente de transfert vers le second module 20.

Dans la figure 2 représentant un schéma fonctionnel du second module 20 dit de prétraitement des textes, sur commande d'un premier signal H émis par les seconds moyens d'horloge 21, un ensemble texte et cellule associée est transféré dans des moyens de segmentation 22 destinés à séparer les différentes phrases constituant ledit texte court reçu.

Pour ce faire, un indicateur de fin de phrase est inséré après chaque signal de fin de phrase (".", "?", "!", ";") suivi d'un espace. Cet indicateur de fin de phrase est enlevé si le caractère suivant est en minuscule traduisant le fait que la phrase continue. Certaines abréviations sont prises en compte dans l'analyse pour éviter une séparation erronée de phrase.

L'indicateur de fin de phrase peut être constitué du numéro du premier caractère de la phrase dans le texte suivi du numéro du dernier caractère de cette phrase, ces numéros étant obtenus par des premiers moyens de compteur 22' connectés aux moyens de segmentation 22.

Par exemple, le texte : " –Pr. Dupont, vous êtes chercheur en linguistique computationnelle,c-à-d. en linguistique informatique. Quelle est la part réelle de linguistique dans votre travail ?"
devient : " –Pr. Dupont, vous êtres chercheur en linguistique computationnelle,c-à-d. en linguistique informatique. [0-101]. Quelle est la part réelle de linguistique dans votre travail ? [103 – 165]".

Les premiers moyens de compteur 22' peuvent également compter les phrases et les numéroter dans l'ordre du texte, le numéro attribué à une phrase jouant un rôle d'index de façon à transférer, phrase par phrase et dans l'ordre croissant des index, les phrases segmentées dans une première file d'attente 23 de type premier entré-premier sorti (FIFO) ou, de préférence, avec clé de tri et priorité.

Sur commande d'un second signal H délivré par lesdits seconds moyens d'horloge 21, une phrase de ladite file d'attente 23 est introduite dans des moyens de découpage 24, connus également sous le nom de tokénisation, destinés à découper une phrase donnée en mots ou, plus généralement, en unités lexicales élémentaires par insertion d'indicateurs de fin de mot après chaque espace reconnu entre les mots de la phrase. Cet indicateur est constitué des numéros de début et de fin de mot comptés dans le texte court complet. Il est déterminé à partir de seconds moyens de compteur 24' éventuellement associés aux premiers moyens de compteur 22', et pouvant indexer chaque mot par le numéro du mot dans la phrase.

L'index résultant permet de transférer, mot par mot, et dans l'ordre des index, les mots reconnus dans une seconde file d'attente 25 de type premier entré-premier sorti (FIFO) ou, de préférence, avec clé de tri et priorité.

Un troisième signal de commande H issu desdits seconds moyens d'horloge 21 autorise le transfert des mots d'une phrase stockés dans la seconde file d'attente 25, mot après mot, vers des moyens d'analyse morphologique et morpho-syntaxique 26.

Dans le moteur linguistique XeLDA® de la Société TEMIS, déjà cité ci-dessus, des automates déterministes d'état fini associés à des lexiques mettent en oeuvre une analyse morphologique et morpho-syntaxique destinée à reconnaître pour chaque mot la forme normalisée (lemme en forme de base) et toutes les catégories

grammaticales exactes d'un mot en fonction de son contexte proche, c'est à dire en fonction d'un ou deux mot(s) avant lui ou après lui.

Par ces moyens, il est possible également de traiter les problèmes liés à la reconnaissance de mots en majuscule, de mots mal accentués, de certaines abréviations, etc.

La levée de certaines ambiguïtés est résolue à l'aide desdits moyens 26 comme, par exemple, le mot d'une phrase "avions" qui peut porter deux sens, l'un provenant du verbe avoir (nous avions) et l'autre étant le nom avions au pluriel.

Pour mieux illustrer les différentes étapes mises en oeuvre dans les différents modules, il est possible de considérer l'exemple suivant de texte court (une phrase) écrit en français et concernant le domaine de l'immobilier : "Vds Nantes Centre 200 keuros FAC F2 Récent 38 m^2 chambre, kitch, parkg, interph, Tel 06 84 20 68 04".

Une séparation des mots par les moyens de découpage 24 conduit à la suite de mots reconnus, introduite dans la seconde file d'attente 25 et indiquée dans la première colonne du tableau 1 ci-dessous.

TABLEAU 1

Après découpage (24)	Après analyse (26)	→	Fiche d'Analyse
MOT	MOT	ET	ATTRIBUT (S)
Vds	Vendre		verbe
Nantes	Nantes		nom-propre
Centre	Centre		nom-singulier
200	200		nombre
keuros	keuro		unité
FAC	FAC		mot inconnu
F2	F2		mot inconnu
Récent	Récent		adjectif
38	38		nombre
m^2	m^2		unité

Chambre	chambre	nom-singulier
kitch	kitch	mot inconnu
parkg	parkg	mot inconnu
Tel	Tel	mot inconnu
0684206804	0684206804	nombre

La seconde colonne du tableau 1 intitulée "MOT ET ATTRIBUT(S)" présente les résultats de l'analyse réalisée par les moyens 26 qui associent à chaque mot un ou plusieurs attributs représentant les catégories grammaticales identifiées.

Cette seconde colonne qui constitue une fiche d'analyse (FA) est d'abord stockée dans une troisième file d'attente FIFO 27 ou, de préférence, avec clé de tri et priorité, avant d'être mémorisée sur commande d'un quatrième signal d'horloge H dans des seconds moyens de mémoire 28 à des fins d'archivage intermédiaire, d'études ultérieures ou de reprise du processus vers le troisième module 30 décrit ci-dessous en relation avec la figure 3.

La figure 3 représente un schéma fonctionnel du troisième module 30 dit d'extraction des informations.

L'élément fondamental de ce module 30 est le dispositif d'extraction 31 constitué d'une pluralité d'automates et de transducteurs d'état fini résultant d'une compilation de données lexicales et de règles d'extraction. Ces données lexicales et ces règles d'extraction sont déterminées pour un domaine spécifique. Le système devra donc posséder au moins un tel dispositif d'extraction par domaine susceptible d'être concerné par le texte court reçu et que le gestionnaire du serveur souhaite pouvoir traiter.

Un serveur d'extraction d'information approprié pour s'adapter au mode préféré de réalisation de l'invention concerne l'outil mis au point par la Société TEMIS, intitulé Insight Discover™ Extractor.

L'extraction de l'information se fait à partir des mots et de leurs attributs associés rassemblés dans ladite fiche d'analyse (FA), l'identification de l'information portée par chacun desdits mots et des attributs associés se traduisant par la définition de nouveaux attributs et convertissant ainsi chaque fiche d'analyse en une nouvelle fiche désignée par "fiche préliminaire" FP.

Cette conversion est illustrée dans le Tableau 2 ci-dessous qui représente une fiche préliminaire établie dans le cas de l'exemple considéré précédemment.

TABLEAU 2 - Fiche Préliminaire

MOTS		NOUVEAUX ATTRIBUTS
Vendre	=	\ Objet \ Vendre
Nantes	=	\ Ville \ Nantes
Centre	=	\ Localisation \ Centre ville
200 keuro	=	\ Prix \ 200 keuro
FAC	=	\ Frais d'agence \ inclus
F2	=	\ Type \ F2
Récent	=	\ Période de construction \ récent
38 m^2	=	\ Surface \ 38 m^2
Chambre	=	\ Nombre de Chambre \ 1
kitch	=	\ kitchenette \ oui
parkg	=	\ parking \ oui
tel 0684206804	=	\ téléphone \ portable \ 0684206804

Chaque fiche préliminaire est stockée dans des troisièmes moyens de mémoire 32 qui peuvent être constitués d'une mémoire associative dont les positions de mémoire sont identifiées par leur contenu, la clé de recherche associée à chaque donnée enregistrée étant le mot binaire de la cellule Cti caractéristique de la $i^{ème}$ phrase du texte correspondant lui-même à la cellule Ct, cette clé Cti restant invariante à travers les processus décrits ci-dessus.

A partir d'un dispositif d'extraction spécifique d'un domaine, il est possible de constituer une fiche type (FT) sous forme d'une base de données élémentaire vide d'enregistrement et ne contenant que des noms de champ qui sont de deux types : un premier type de champ pour définir l'objet du texte et une pluralité d'autres champs de second type définissant les paramètres liés à cet objet.

Pour reprendre l'exemple de l'immobilier déjà considéré, on peut définir le champ objet qui sera Vendre, on peut définir dans les champs suivants tous

les paramètres qui peuvent se rapporter au produit en vente, à savoir, lieu, prix, surface, type, nombre de pièces, présence de parking, de piscine, de balcon, etc....

Cette fiche type constitue une base de données élémentaire appartenant à un système de bases de données 33 rassemblant des fiches types provenant d'un ou de plusieurs dispositifs d'extraction 31', 31" et pouvant rassembler des fiches appartenant à des domaines différents mais pouvant présenter certains noms de champ communs.

Enfin, la construction d'une quatrième fiche dite "fiche caractéristique" (FC) permettra d'exploiter le système décrit. Cette fiche caractéristique résulte de la création d'un enregistrement dans la fiche type en remplissant les champs qui se retrouvent dans une fiche préliminaire FP.

En considérant l'exemple déjà mentionné, le tableau 3 ci-dessous décrit la fiche caractéristique créée en utilisant les moyens de chargement 34 des données d'une fiche préliminaire dans une fiche type (FT) appropriée présente dans la mémoire vive du système de bases de données 33.

TABLEAU 3 – Fiche Caractéristique

OBJET :		=	Vendre
PARAMETRES :	Ville	=	Nantes
	Localisation	=	Centre Ville
	Prix	=	200 keuro
	Frais d'agence	=	inclus
	Type	=	F2
	Période de construction	=	Récente
	Surface	=	38 m^2
	Nombre de chambre	=	1
	Kitchenette	=	oui
	Parking	=	oui
	Téléphone portable	=	0684206804

Des quatrièmes moyens de file d'attente avec priorité 35 stockent temporairement des fiches caractéristiques et les délivrent, soit à des quatrièmes

moyens de mémoire 36 constitués comme les troisièmes moyens de mémoire 32 d'une mémoire associative avec clé de recherche représentée par la cellule Cti, soit au quatrième module 40.

Des troisièmes moyens d'horloge 37 sont également prévus pour émettre différents signaux H' de commande d'accès aux différents composants du module 30.

La figure 4 représente un schéma fonctionnel du quatrième module 40 dit d'exploitation et de diffusion ciblée.

Des quatrièmes moyens de mémoire 36 ou des quatrièmes moyens de file d'attente avec priorité 35, est extraite une fiche caractéristique FCi dont l'ensemble des champs est introduit en parallèle dans un réseau à logique programmable 41. Ce réseau 41 est programmé pour comparer la fiche FCi avec un ensemble de fiches caractéristiques présélectionnées et extraites de la mémoire 36 par les moyens de sélection 42 et stockées finalement dans une table à consulter 43 qui fonctionne avec ledit réseau 41.

Lorsqu'une fiche caractéristique présélectionnée présente des champs de même valeur que ceux de la fiche FCi ou vérifient une relation préétablie et introduite dans la programmation du réseau 41, une des sorties du réseau 41 fournit les données de la fiche caractéristique présélectionnée en coïncidence avec la fiche FCi, ces données étant alors stockées dans des seconds moyens de registre 44.

Le nombre de coïncidences est obtenu à partir d'un compteur 47 qui peut indiquer, ainsi, le nombre de fiches qui répondent positivement à une comparaison avec la fiche FCi.

Des moyens de balayage 45 suivent l'ensemble des registres 44 et transfèrent chaque fiche caractéristique identifiée comme répondant positivement dans des cinquièmes moyens de file d'attente avec priorité 46 à des fins de stockage temporaire avant, soit de les mémoriser dans des cinquièmes moyens de mémoire 49 constitués, par exemple, d'une mémoire associative comme les moyens de mémoire 32, soit de les diffuser par les moyens de diffusion 48.

Ces derniers équipés de moyens de décodage 48' et de conversion 48" peuvent transmettre à l'utilisateur émetteur du texte court, ou à tous autres destinataires autorisés, en langage clair, le contenu des fiches caractéristiques

identifiées comme des réponses positives. Cette diffusion est correctement ciblée car la cellule Cti associée, récupérée dans les mémoires 49 ou 36, indique les caractéristiques de l'utilisateur émetteur et le support de communications à sélectionner.

Les moyens 48, 48', 48" peuvent être regroupés avec les moyens 7, 7', 7".

Des quatrièmes moyens d'horloge 50 permettent d'horodater la transmission des réponses à l'utilisateur et d'émettre, comme pour les modules précédents, des signaux H" de commande des accès à différents composants du quatrième module 40.

En poursuivant la description de l'exemple considéré précédemment, il est possible d'illustrer un avantage supplémentaire du système et du procédé selon l'invention.

En effet, dans l'application au traitement automatique des petites annonces dont le contenu concerne des demandes ou des offres de produits ou de services, l'exemple considéré susdit est très représentatif du domaine de l'immobilier. Le tableau 3 indique que l'objet de l'annonce est une Vente (vendre), c'est à dire une offre, et l'objet complémentaire, c'est à dire une demande, sera un Achat (acheter). Les autres champs du tableau 3 représentent les paramètres du produit à vendre.

Le procédé consiste à :

présenter la fiche caractéristique (FCi) Vente à l'entrée du réseau à logique programmable 41,

sélectionner 42 dans la mémoire 36 toutes les fiches caractéristiques dont l'objet est complémentaire, à savoir des fiches d'objet = Achat qui seront utilisées à des fins de comparaison par l'intermédiaire du réseau 41 fonctionnant en liaison avec la table à consulter 43,

comparer dans ledit réseau 41 tous les champs de paramètres des fiches Achat avec ceux de la fiche Vente,

stocker en sortie du réseau 41 les fiches Achat dont les valeurs des paramètres coïncident avec toutes les valeurs des paramètres de la fiche Vente, les fiches Achat ainsi identifiées constituant les réponses positives à la comparaison

seront à diffuser à l'utilisateur émetteur ou à tous autres destinataires autorisés de la fiche Vente sur le support de communications approprié.

A partir du dispositif d'extraction 31 spécifique du domaine immobilier, il est possible d'extraire une fiche type (FT) contenant tous les champs de paramètres imaginables pouvant décrire et qualifier un produit immobilier.

Une telle fiche type sera utilisée dans deux cas particuliers.

1/ La comparaison conduit à un nombre de réponses positives trop élevé, c'est à dire qu'il devient difficile pour le vendeur de faire le tri dans, par exemple, une centaine de demandes d'achat. Dans ce cas, on transmet à l'annonceur utilisateur du système, par les moyens 48, la liste des champs de paramètres afin qu'il complète des paramètres non définis dans sa première annonce ou qu'il modifie dans un sens plus restrictif certaines valeurs de paramètre. Ces paramètres modifiés ou supplémentaires créent une nouvelle fiche caractéristique FCi qui comporte, en conséquence, plus de champs de paramètre. Une réponse positive devra ainsi satisfaire à un plus grand nombre de valeurs identiques de paramètres et limitera, de ce fait, le nombre de réponses positives.

Un nombre maximal prédéterminé de réponses peut être introduit dans les moyens de compteur 47 qui, dès l'obtention de ce nombre maximal, arrêteront le processus du réseau 41 et activeront simultanément une commande de transmission d'une fiche type vers l'annonceur.

2/ La comparaison conduit à l'absence de réponse positive. Dans ce cas, on demandera par les mêmes moyens que précédemment à l'utilisateur de supprimer des paramètres ou de modifier dans un sens moins restrictif les valeurs de certains champs de paramètre afin de présenter une nouvelle fiche caractéristique FCi et augmenter ainsi la probabilité de trouver au moins une réponse positive en sortie du réseau à logique programmable.

Les modifications, les ajouts ou les retraits de valeurs de paramètres peuvent faire l'objet de plusieurs étapes sous forme de processus en boucle permettant de rechercher l'adaptation optimale entre l'offre et la demande.

Ce procédé de traitement des petites annonces est très bien adapté à l'utilisation des réseaux de téléphonie fixe ou mobile en tant que support de

communications, la transformation en données numériques du texte court étant réalisée, dans ce cas, à l'aide d'un serveur vocal classique.

Plusieurs modifications ou variantes sont susceptibles d'être apportées au mode préféré de réalisation de l'invention décrit ci-dessus, mais elles restent dans le champ de l'invention précisé par les revendications jointes.

REVENDICATIONS

1. Système automatique d'extraction et de traitement des informations portées par des textes courts accompagnés, ou non, de données audio et/ou vidéo et/ou photographiques, transmis sur tous supports de communications et de diffusion ciblée sur tous supports de communications des résultats de recherche de relations entre ces informations ou de comparaisons établies entre elles, caractérisé en ce que ledit système est rendu automatique par la mise en oeuvre d'un serveur équipé d'un microprocesseur qui commande et organise la combinaison des modules suivants, comprenant :

A/ au moins un premier module (1), dit de réception et de précodage, comprenant :

- des moyens de récepteur (2) destinés à recevoir, de façon continue ou non, une pluralité de textes courts transmis sur un quelconque support de communications, chacun d'entre eux étant associé à un premier code (code 1) identifiant l'utilisateur émetteur d'un texte court, et codé en binaire par des premiers moyens de codeur (4')

- des seconds moyens de codeur binaire (2') destinés à identifier sous forme codée (code 2) la nature du support de communications acheminant ledit texte court,

- des moyens de numérisation (A/N) connectés aux moyens de récepteur et destinés à numériser les contenus desdits textes courts non préalablement numérisés par le mode de transmission,

- des premiers moyens d'horloge (3) associés à des troisièmes moyens de codeur binaire (3') destinés à coder (code 3) des données d'horodatage concernant la réception desdits textes courts,

- des moyens d'identification et d'authentification (4) de l'émetteur dudit texte court par vérification du premier code (code 1) qui y est associé,

- des moyens de reconnaissance de la langue (5) dans laquelle ledit texte court est écrit, associés à des quatrièmes moyens de codeur binaire (5') pour identifier sous forme codée (code 4) la langue reconnue,

- des moyens de reconnaissance du domaine (6) concerné par le contenu dudit texte court, associés à des cinquièmes moyens de codeur binaire (6') pour identifier, sous forme codée (code 5), le domaine reconnu,

- une pluralité de bases de données (11) connectée respectivement aux moyens d'identification (4), aux moyens de reconnaissance de la langue (5) et du domaine (6),

- des premiers moyens de registre (10) recevant les différents codes (code 1 à code 5), de façon à constituer, sous forme d'un mot binaire, une cellule (Ct) établie pour être associée en permanence à un et un seul texte court reçu,

- des moyens de concaténation (12) de chaînes binaires destinés à ajouter la cellule (Ct) à la chaîne de sortie des moyens de reconnaissance de domaine (6), pour obtenir une chaîne unique regroupant texte court et cellule (Ct),

- des moyens de transmission (7), en retour vers l'utilisateur émetteur d'origine (13), du texte court reçu par le serveur, converti en langage clair par des moyens de conversion (7") et accompagné de la cellule (Ct) associée, cette dernière étant décodée dans des moyens de décodeur (7'),

- des moyens de validation (13') dépendant du support de communications utilisé et destinés à délivrer un signal de validation (V) généré par ledit utilisateur émetteur de texte court en réponse à son approbation du texte et de la cellule (Ct) qui lui ont été retransmis, toute modification de la cellule (Ct) devenant impossible après ladite validation,

- des moyens de circuit logique (8) recevant en première entrée la chaîne unique produite par les moyens de concaténation (12) et ne pouvant délivrer en sortie cette même chaîne que si le signal de validation (V) est appliqué en seconde entrée,

- des premiers moyens de mémoire (9) destinés à stocker ladite chaîne unique produite par les moyens de concaténation (12) regroupant texte court et cellule (Ct), l'accès aux dits premiers moyens de mémoire étant commandés par lesdits moyens de circuit logique (8) ;

B/ au moins un second module (20), dit de prétraitement des textes, comprenant:

- des moyens de segmentation (22) destinés à séparer les phrases formant ledit texte court et associés à des premiers moyens de compteur (22') réservés à la

numérotation des différentes phrases reconnues et à la numérotation de chaque caractère utilisé pour former ces phrases dudit texte court,

- des moyens de découpage (24) destinés à séparer les mots d'une phrase reconnue, et associés à des seconds moyens de compteur (24') réservés à la numérotation attribuée à chacun des caractères de début et de fin de chaque mot par rapport à l'ensemble des mots identifiés dans une phrase reconnue,

- des moyens d'analyse morphologique et morpho-syntaxique (26) de tous les mots identifiés d'une phrase, destinés à transformer chaque mot identifié en sa forme normalisée (lemme) et à lui associer des attributs liés à sa catégorie grammaticale exacte en fonction de son contexte proche,

- des premiers moyens de file d'attente (23) destinés à stocker les différentes phrases identifiées et numérotées en sortie desdits moyens de segmentation (22),

- des seconds moyens de file d'attente (25) destinés à stocker les différents mots identifiés dont les caractères ont été numérotés en sortie desdits moyens de découpage (24),

- des troisièmes moyens de file d'attente (27) destinés à stocker les mots identifiés normalisés et leurs attributs associés en sortie desdits moyens d'analyse morphologique et morpho-syntaxique (26),

- des seconds moyens de mémoire (28) destinés à stocker en sortie des troisièmes moyens de file d'attente (27) une pluralité de fiches d'analyse (FA) constituées, chacune, des mots identifiés normalisés et de leurs attributs associés correspondant à une phrase donnée, chaque fiche d'analyse (FA) étant accompagnée de la cellule (Ct) correspondant audit texte court d'origine, cette dernière étant invariante au cours des traitements cités,

- des seconds moyens d'horloge (21) destinés à délivrer, selon des séquences pré-programmées, des signaux de commande d'accès (H), respectivement, aux différents moyens de traitement (22, 24, 26) et de mémoire (28) ;

C/ au moins un troisième module (30), dit d'extraction des informations, comprenant :

- une pluralité de dispositifs d'extraction des informations (31), chacun étant spécifique d'un domaine, ces dispositifs (31) étant destinés à transformer les

attributs associés à chaque mot de la fiche d'analyse (FA) en de nouveaux attributs identifiant l'information portée par chaque mot, convertissant ainsi chaque fiche d'analyse (FA) en une fiche préliminaire (FP),

- des troisièmes moyens de mémoire (32) destinés à stocker lesdites fiches préliminaires (FP) et leur cellule (Ct) associée qui reste invariante en traversant un dispositif d'extraction (31) quelconque,

- au moins un système de bases de données (33) comportant, pour chaque domaine spécifique traité par un dispositif d'extraction (31) donné, une base élémentaire de données, vide d'enregistrement, constituée uniquement d'une pluralité de noms de champ qui définit une fiche type (FT), le système de bases de données (33) comportant également son système de gestion et son dispositif de mémorisation,

- des moyens de chargement (34) des données contenues dans une fiche préliminaire (FP), dans ladite base élémentaire de données, les nouveaux attributs de ladite fiche préliminaire (FP) correspondant aux noms de champ de la fiche type (FT) associée, le résultat étant l'établissement d'une fiche caractéristique (FC) organisée de façon prédéterminée et destinée à identifier l'objet et ses paramètres associés de l'information portée par chaque phrase d'un texte court reçu,

- des quatrièmes moyens de file d'attente (35) destinés à stocker de façon temporaire lesdites fiches caractéristiques (FC) issues dudit système de bases de données (33),

- des quatrièmes moyens de mémoire (36) connectés aux quatrièmes moyens de file d'attente (35) et destinés à stocker de façon permanente lesdites fiches caractéristiques (FC) et les cellules (Ct) associées,

- des troisièmes moyens d'horloge (37) destinés à délivrer, selon une séquence pré-programmée, des signaux de commande (H') des accès, respectivement, à la pluralité de dispositifs d'extraction (31), audit système de bases de données et aux dits quatrièmes moyens de mémoire (36) ;

D/ au moins un quatrième module (40), dit d'exploitation et de diffusion ciblée, comprenant :

- des moyens de sélection (42), connectés aux quatrièmes moyens de mémoire (36) destinés à sélectionner une ou plusieurs fiche(s) caractéristique(s) (FC) et à les transférer dans une table à consulter (43),

- au moins un réseau à logique programmable (41) destiné à comparer sur des critères préprogrammés une fiche caractéristique courante (FCi) provenant, soit des quatrièmes moyens de file d'attente (35), soit des quatrièmes moyens de mémoire (36) avec une ou plusieurs autre(s) fiche(s) caractéristique(s) (FC) présente(s) dans ladite table à consulter (43) et à délivrer en sortie le contenu de la (ou des) fiche(s) caractéristique(s) identifiée(s) comme répondant positivement à ladite comparaison,

- des seconds moyens de registre (44) destinés, chacun, à stocker temporairement le contenu d'une fiche caractéristique (FC) identifiée comme répondant positivement à ladite comparaison et délivrée par ledit réseau à logique programmable (41),

- des moyens de balayage (45) destinés à analyser successivement les sorties respectives desdits moyens de registre (44) et à transférer les contenus de ces derniers (44),

- des cinquièmes moyens de file d'attente (46) destinés à stocker temporairement les fiches caractéristiques (FC) identifiées positivement provenant desdits moyens de balayage (45),

- des troisièmes moyens de compteur (47) connectés aux dits seconds moyens de registre (44) et destinés à compter le nombre de fiches caractéristiques (FC) identifiées comme répondant positivement à la comparaison exécutée par ledit réseau à logique programmable (41),

- des cinquièmes moyens de mémoire (49) destinés à stocker à des fins de sauvegarde, d'archivage ou d'études ultérieures, les fiches caractéristiques (FC) identifiées délivrées par les cinquièmes moyens de file d'attente (46) ainsi que les cellules (Ct) associées,

- des moyens de diffusion (48) destinés à une diffusion ciblée sur le support de communication approprié des fiches caractéristiques identifiées positivement provenant, dans un ordre préprogrammé, des cinquièmes moyens de file d'attente (46) ou des cinquièmes moyens de mémoire (49), ces moyens de diffusion (48) étant équipés de moyens de décodage (48') et de conversion (48") des données numériques, ces derniers moyens (48', 48") s'appliquant au contenu des fiches caractéristiques (FC) identifiées positivement et aux cellules (Ct) associées,

- des quatrièmes moyens d'horloge (50) destinés à horodater la diffusion ciblée exécutée par lesdits moyens de diffusion (48) et à délivrer des signaux de commande (H") d'accès, respectivement, audit réseau à logique programmable (41), aux dits moyens de balayage (45), aux dits moyens de diffusion (48) et aux dits cinquièmes moyens de mémoire (49).

2. Système selon la revendication 1, caractérisé, de plus, en ce que lesdits moyens de reconnaissance de la langue (5) mettent en oeuvre un procédé à base de trigrammes et de mots courts.

3. Système selon la revendication 1 ou 2, caractérisé, de plus, en ce que lesdits moyens de reconnaissance du domaine (6) mettent en oeuvre un procédé de catégorisation avec apprentissage.

4. Système selon l'une quelconque des revendications précédentes caractérisé, de plus, en ce que le dispositif d'extraction est constitué d'automates et de transducteurs informatiques d'état fini résultant d'une compilation de données lexicales et de règles d'extraction.

5. Système selon l'une quelconque des revendications précédentes, caractérisé, de plus, en ce que les textes courts, accompagnés ou non de données audio et/ou vidéo et/ou photographiques, sont constitués d'annonces dont le contenu concerne des demandes ou des offres de produits et/ou de services, et en ce que le support de communications est un réseau de téléphonie fixe ou mobile connecté à un serveur vocal classique en tant que moyens de numérisation (A/N).

6. Procédé automatique d'extraction et de traitement des informations portées par des textes courts mettant en oeuvre le système selon l'une quelconque des revendications précédentes, caractérisé par les étapes suivantes consistant à :

a) entrer en communication avec le module de réception (1) d'un serveur auquel sont transmis un texte court accompagné ou non de données audio et/ou vidéo et/ou photo et un code utilisateur (code 1), ledit module (1) numérisant (A/N), si nécessaire,

les données reçues, vérifiant (4) l'identification de l'utilisateur, et horodatant (3) la réception dudit texte court ;

b) déterminer la langue (5) et le domaine (6) concernés par ledit texte court reçu;

c) faire valider par l'utilisateur le contenu du texte reçu, la langue et le domaine reconnus ;

d) séparer les phrases (22) du texte et les mots (24) de chacune des phrases ;

e) pratiquer (26) sur chaque mot une analyse morphologique et morpho-syntaxique dont le résultat est de produire un mot normalisé auquel est associé un attribut (FA) ;

f) extraire à partir de moyens d'extraction (31) spécifiques du domaine reconnu à l'étape b l'information portée par chaque mot associé à son attribut pour convertir ce dernier en un ou plusieurs nouveau(x) attribut(s) parmi lesquels sont reconnus l'objet dudit texte court et les paramètres liés à cet objet ;

g) mémoriser (32) chaque mot d'une phrase dudit texte court reçu avec ses nouveaux attributs dans une fiche préliminaire (FP) ;

h) obtenir une fiche caractéristique (FC) (33), en créant (34) un enregistrement dans une fiche type (FT) constituée d'un champ objet et d'une pluralité de champs de paramètres à partir des données contenues dans une fiche préliminaire (FP) ;

i) mémoriser (36) l'ensemble des fiches caractéristiques (FC) ;

j) comparer (41) une ou plusieurs fiche(s) caractéristique(s) (FCi) à une ou plusieurs fiche(s) caractéristique(s) sélectionnée(s) (42, 43) ;

k) sortir (44, 45) les résultats de chaque comparaison effectuée et les mémoriser (49) ;

l) diffuser (48) les résultats mémorisés et convertis (48") vers l'utilisateur émetteur dudit texte court reçu ou vers tous autres destinataires autorisés.

7. Procédé automatique selon la revendication 6, caractérisé, de plus, par les étapes supplémentaires consistant à coder en binaire (2', 3', 4', 5', 6'), respectivement, le type de support de communications utilisé pour la transmission dudit texte court reçu, les données d'horodatage de la réception de ce dernier, le code utilisateur, la langue reconnue et le domaine reconnu concernant ledit texte court reçu, pour constituer un mot binaire contenu dans une cellule (Ct, Cti) qui restera associée à un texte court

reçu, cette cellule n'étant plus accessible en écriture après validation (V) par l'utilisateur émetteur dudit texte court.

8. Procédé automatique selon la revendication 6 ou 7, caractérisé, de plus, par la mise en oeuvre de signaux (H, H', H") d'horloge (3, 21, 37) destinés à commander les accès aux différents moyens de traitement (22, 24, 26, 31, 33, 41, 42, 45), de mémorisation (28, 32, 36, 49) et de diffusion (48) du système.

9. Procédé automatique d'extraction et de traitement des informations portées par des textes d'annonces dont le contenu concerne des demandes ou des offres de produits et/ou de services mettant en oeuvre le système selon l'une quelconque des revendications 1 à 5 et le procédé selon l'une quelconque des revendication 6 à 8, caractérisé en ce que les comparaisons réalisées (41) entre fiches caractéristiques (FC) portent uniquement sur les champs de paramètres de fiches caractéristiques (FC) dont les objets sont complémentaires.

10. Procédé selon la revendication 9, caractérisé, de plus, en ce qu'il comporte :

- en cas d'un nombre élevé de réponses positives (47) en sortie des moyens de comparaison (41), correspondant, soit à une offre, soit à une demande, une ou plusieurs étape(s) de limitation de ce nombre en modifiant les valeurs de certains champs de paramètre dans un sens plus restrictif ou en faisant porter les comparaisons sur un plus grand nombre de champs de paramètre, ou les deux,

- en cas d'absence de réponse positive en sortie des moyens de comparaison (41) correspondant, soit à une offre, soit à une demande, une ou plusieurs étape(s) de modification des valeurs de certains champs de paramètre dans un sens moins restrictif ou de suppression d'un ou de plusieurs champs de paramètres à prendre en compte dans la comparaison (41) entre fiches caractéristiques, ou les deux ,

dans les deux cas, lesdites étapes mettent en oeuvre des échanges (48) entre système et utilisateur de façon à déterminer de façon optimale les noms de champs de paramètre à prendre en compte en plus ou en moins dans les comparaisons (41) entre fiches caractéristiques.

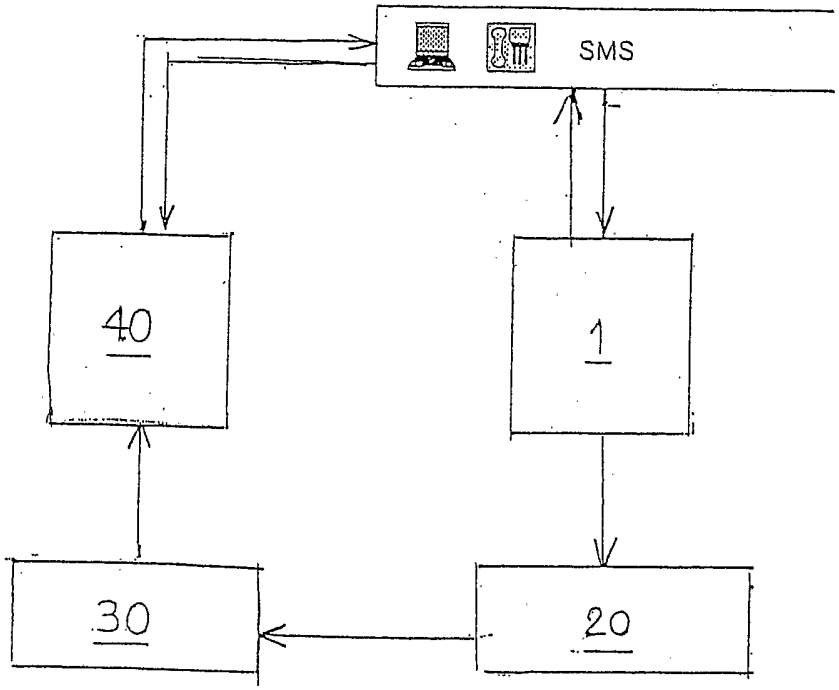


Figure 1

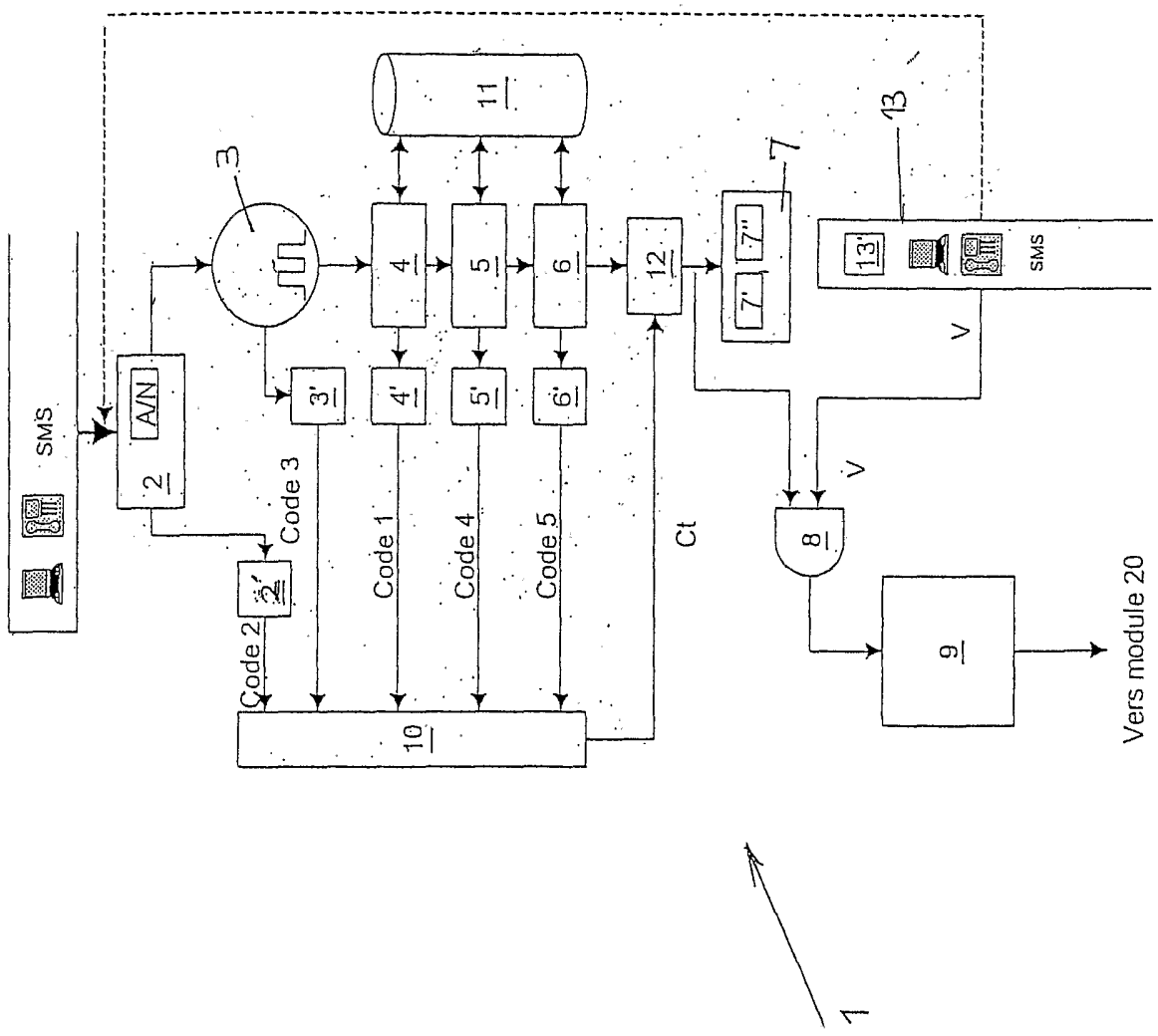

SMS
A/N
2
3
11
4
5
6
12
7
7'
7"
13
13'
SMS
V
3'
4'
5'
6'
Code 3
Code 1
Code 4
Code 5
Code 2
2'
Ct
8
V
9
Vers module 20
10
1

Figure 2

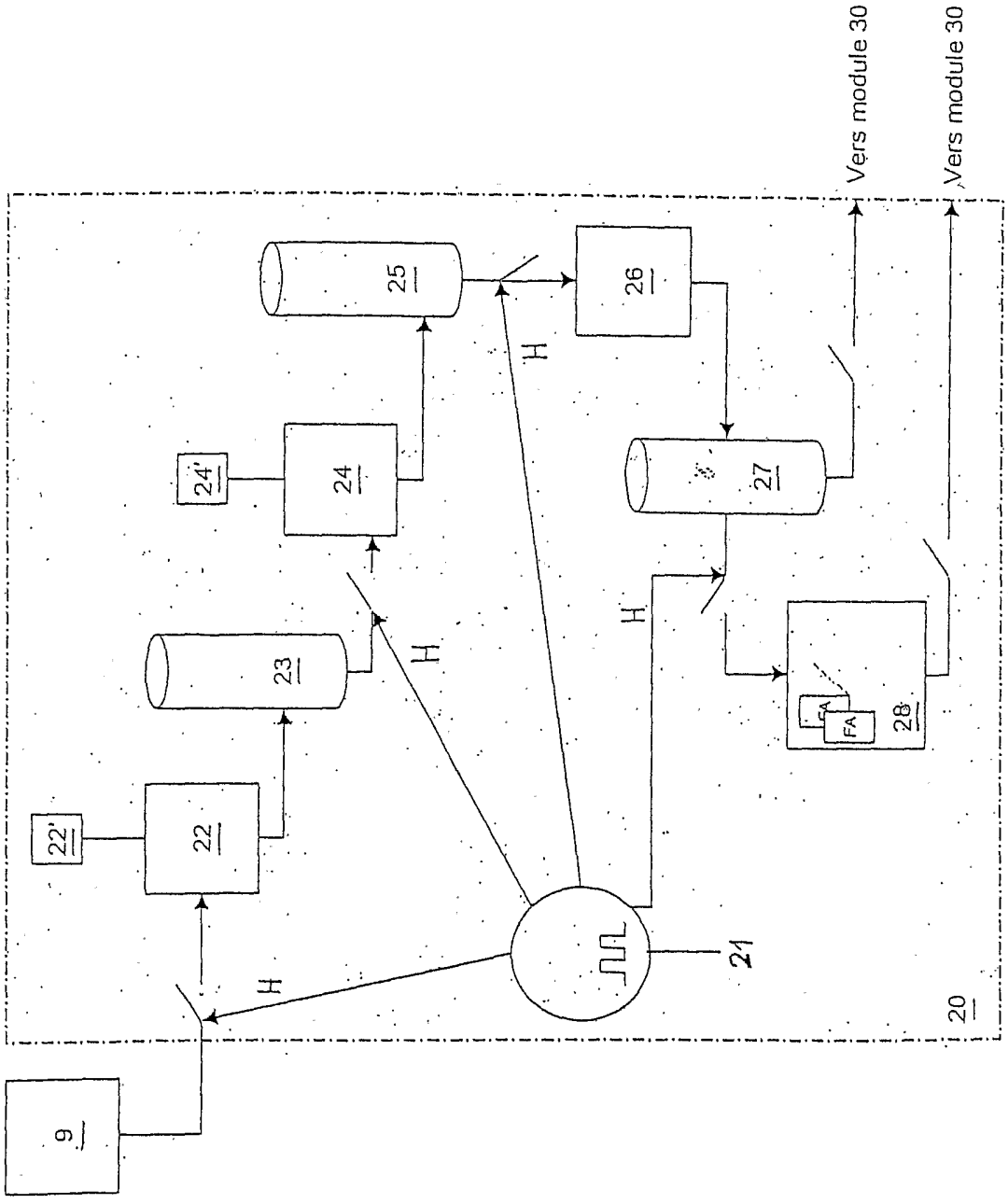

Vers module 30
Vers module 30
25
26
24'
24
27
23
22'
22
28
FA
FA
H
H
H
H
21
20
9

Figure 3

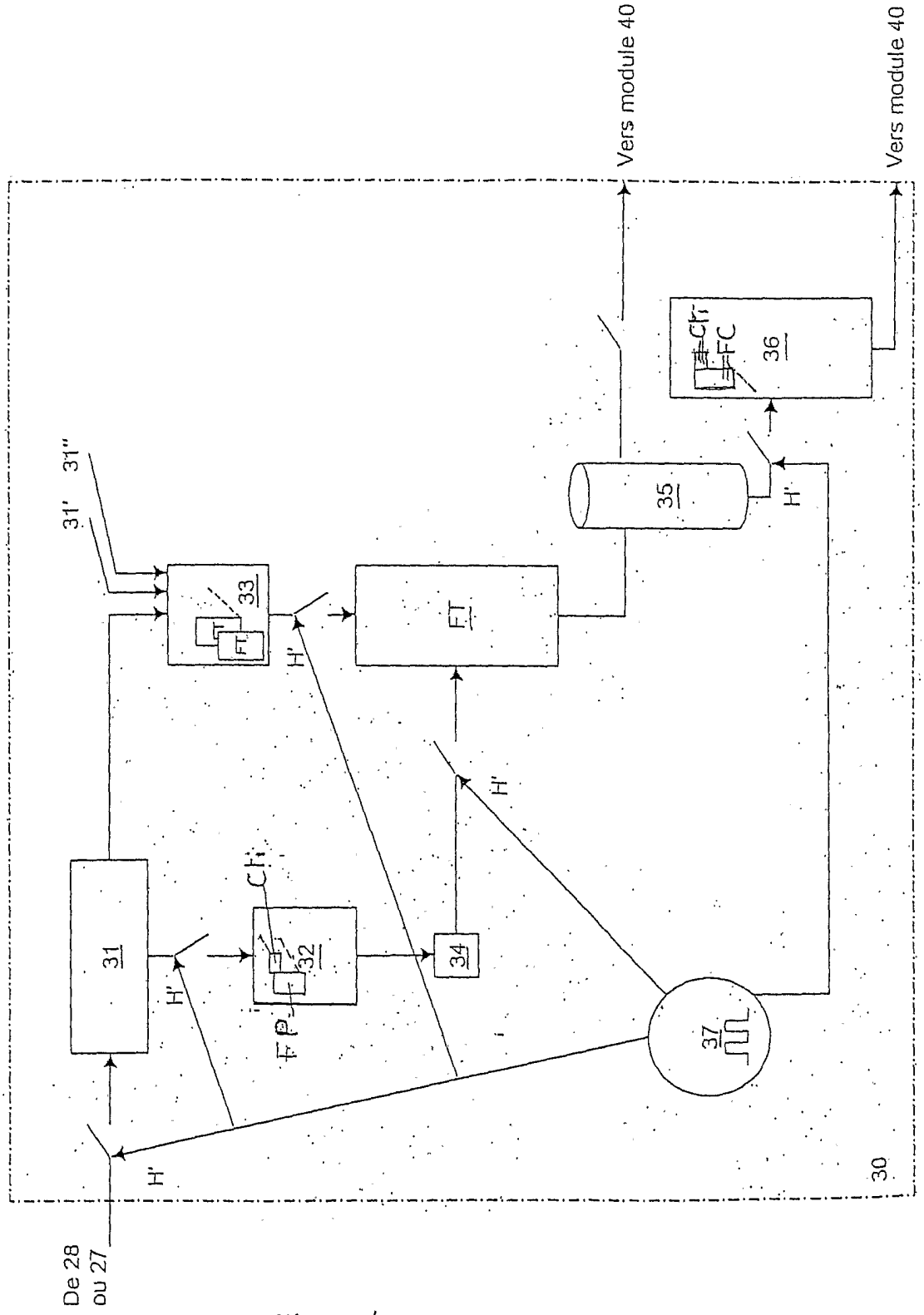

Vers module 40
Vers module 40
31' 31''
33
FT
H'
FT
35
H'
36
FC
H'
31
H'
32
FP
34
37
H'
30
De 28
ou 27

Figure 4

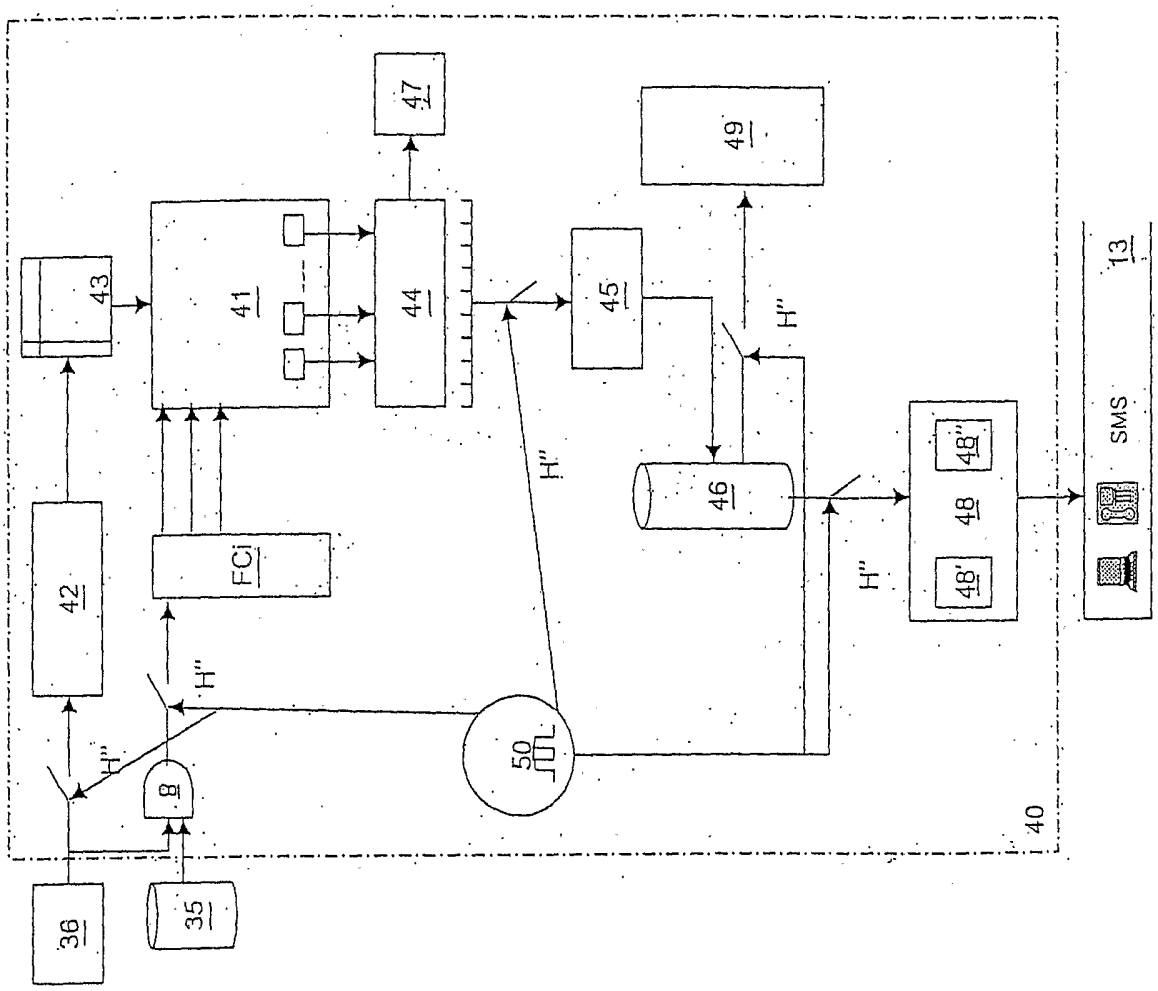


Figure 5

INTERNATIONAL SEARCH REPORT

Intern ıal Application No
PCT/FR2005/000023

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G06F17/30

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 G06F H04Q

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

EPO-Internal

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	WO 01/08430 A (INTELLPROP LTD; WILSON JEFFREY) 1 February 2001 (2001-02-01) abstract page 2, line 17 - page 5, line 31 figure 1	1-10
A	EP 1 014 276 A (XEROX CORP) 28 June 2000 (2000-06-28) abstract column 14, line 5 - column 18, line 3	1-10
A	KARTTUNEN L: "CONTRUCTING LEXICAL TRANSDUCERS" PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS, XX, XX, 1994, pages 406-411, XP000789867 the whole document	1-10

☐ Further documents are listed in the continuation of box C.

☒ Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
3 May 2005	31/05/2005

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Dumitrescu, C

INTERNATIONAL SEARCH REPORT

ormation on patent family members

Intern al Application No
PCT/FR2005/000023

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 0108430 A	01-02-2001	AU 6300300 A	13-02-2001
		WO 0108430 A1	01-02-2001
EP 1014276 A	28-06-2000	US 6167369 A	26-12-2000
		BR 9905978 A	05-09-2000
		EP 1014276 A2	28-06-2000
		JP 2000194696 A	14-07-2000

RAPPORT DE RECHERCHE INTERNATIONALE

Dema nternationale No
PCT/FR2005/000023

A. CLASSEMENT DE L'OBJET DE LA DEMANDE
CIB 7 G06F17/30

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

B. DOMAINES SUR LESQUELS LA RECHERCHE A PORTE

Documentation minimale consultée (système de classification suivi des symboles de classement)
CIB 7 G06F H04Q

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

Base de données électronique consultée au cours de la recherche internationale (nom de la base de données, et si réalisable, termes de recherche utilisés)
EPO-Internal

C. DOCUMENTS CONSIDERES COMME PERTINENTS

Catégorie °	Identification des documents cités, avec, le cas échéant, l'indication des passages pertinents	no. des revendications visées
A	WO 01/08430 A (INTELLPROP LTD; WILSON JEFFREY) 1 février 2001 (2001-02-01) abrégé page 2, ligne 17 - page 5, ligne 31 figure 1	1-10
A	EP 1 014 276 A (XEROX CORP) 28 juin 2000 (2000-06-28) abrégé colonne 14, ligne 5 - colonne 18, ligne 3	1-10
A	KARTTUNEN L: "CONTRUCTING LEXICAL TRANSDUCERS" PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS, XX, XX, 1994, pages 406-411, XP000789867 le document en entier	1-10

☐ Voir la suite du cadre C pour la fin de la liste des documents

☒ Les documents de familles de brevets sont indiqués en annexe

° Catégories spéciales de documents cités:

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

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

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

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

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

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

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

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

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

Date à laquelle la recherche internationale a été effectivement achevée	Date d'expédition du présent rapport de recherche internationale
3 mai 2005	31/05/2005

Nom et adresse postale de l'administration chargée de la recherche internationale	Fonctionnaire autorisé
Office Européen des Brevets, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31–70) 340–2040, Tx. 31 651 epo nl, Fax: (+31–70) 340–3016	Dumitrescu, C

RAPPORT DE RECHERCHE INTERNATIONALE

Renseignements relatifs à nembres de familles de brevets

Dem Internationale No
PCT/FR2005/000023

Document brevet cité au rapport de recherche		Date de publication	Membre(s) de la famille de brevet(s)		Date de publication
WO 0108430	A	01-02-2001	AU	6300300 A	13-02-2001
			WO	0108430 A1	01-02-2001
EP 1014276	A	28-06-2000	US	6167369 A	26-12-2000
			BR	9905978 A	05-09-2000
			EP	1014276 A2	28-06-2000
			JP	2000194696 A	14-07-2000