Une mesure de similarité contextuelle pour l'aide à la navigation dans un treillis

Saoussen Sakji*, Marie-Aude Aufaure*,**
Géraldine Polaillon*, Bénédicte Le Grand***, Michel Soto***

*Supélec – Computer Science department – plateau du Moulon – 3, rue Joliot Curie
F-91192 Gif sur Yvette Cedex
{Saoussen.Sakji, Marie-Aude.Aufaure, Geraldine.Polaillon}@supelec.fr
** INRIA Paris-Rocquencourt – Domaine de Voluceau – Rocquencourt
B.P. 105 F-78153 Le Chesnay Cedex
Marie-Aude.Aufaure@inria.fr
***Laboratoire d'Informatique de Paris 6 – 104, av du P^{dt} Kennedy – F-75016 Paris
{Benedicte.Le-Grand, Michel.Soto}@lip6.fr

Résumé. La recherche d'information et la navigation dans les pages web s'avèrent complexes du fait du volume croissant des données et de leur manque de structure. La formalisation conceptuelle d'un *contexte* associé à une ontologie rend possible l'amélioration de ce processus.
Nous définissons un *contexte conceptuel* comme étant l'association d'un treillis de concepts construit à partir de pages web avec des ontologies. La recherche et la navigation peuvent alors s'effectuer à plusieurs niveaux d'abstraction : le niveau des données, le niveau conceptuel et le niveau sémantique. Cet article s'intéresse essentiellement au niveau conceptuel grâce à une représentation par les treillis de concepts des documents selon les termes qu'ils ont en commun. Notre objectif est de proposer une mesure de similarité permettant à l'utilisateur de mieux naviguer dans le treillis. En effet, une bonne interprétation du treillis devrait passer par un choix rigoureux des concepts, objets, relations et propriétés les plus intéressants. Pour faciliter la navigation, il faut pouvoir indiquer à l'utilisateur les concepts les plus pertinents par rapport au concept correspondant à sa requête ou pouvoir lui proposer un point de départ. L'originalité de notre proposition réside dans le fait de considérer un lien sémantique entre les concepts du treillis, basé sur une extension des mesures de similarité utilisées dans le cadre des ontologies, afin de permettre une meilleure exploitation de ce treillis. Nous présentons les résultats expérimentaux de l'application de cette mesure sur des treillis construits à partir de pages web dans le domaine du tourisme.

1 Introduction

L'objectif de nos travaux est de faciliter la recherche d'information dans des pages Web par l'utilisation conjointe de treillis de Galois et d'ontologies, qui constitue ce que nous appelons un « contexte conceptuel ». Les regroupements conceptuels fournis par les treillis, associés aux liens sémantiques de l'ontologie, permettent d'améliorer la recherche d'information en fournissant des niveaux de navigation plus abstraits et complémentaires.

Les treillis de Galois restent néanmoins complexes du fait du nombre élevé de concepts qu'ils sont susceptibles de contenir et l'objectif de ce papier est de proposer une mesure de similarité entre ces concepts pour trouver les plus pertinents à conseiller à un utilisateur durant sa navigation dans un treillis.

Ce papier est organisé comme suit : dans un premier temps, la section 2 présente les outils mis en œuvre dans nos travaux, à savoir les treillis de Galois et les ontologies, ainsi que la manière dont ils peuvent être associés pour la recherche d'information. Dans la section 3, nous présentons un état de l'art des mesures de similarité définies pour l'appariement d'ontologies. Nous étendons l'une de ces mesures pour proposer, dans la section 4, une mesure de similarité adaptée aux concepts d'un treillis de Galois ; nous présentons également la manière dont nous la mettons en œuvre pour faciliter la navigation en indiquant les concepts les plus pertinents pour une requête donnée ou à partir de la position courante dans le treillis. Finalement, nous décrivons dans la section 5 une petite expérimentation menée sur un ensemble de pages Web dédiées au tourisme, avant de conclure et de présenter les perspectives de ce travail.

2 Existant

Les travaux présentés ici s'inscrivent dans une méthodologie plus large décrite dans (Le Grand et al., 2006) dont le but est de faciliter la recherche d'information dans des systèmes d'information complexes. Cette méthodologie est fondée sur l'utilisation conjointe de treillis de Galois et de structures sémantiques, par exemple des thesaurus ou des ontologies, pour fournir trois niveaux de navigation dans les données explorées. Ces trois niveaux de navigation proposés sont illustrés sur la figure 1 et décrits dans la section suivante.

2.1 Architecture

La complexité des systèmes d'information peut se traduire de différentes manières, qu'il s'agisse d'un volume important, d'un nombre élevé de dimensions, d'un manque de structure ou des relations et corrélations entre les données du système. Concernant ces deux derniers points en particulier, la construction de treillis de Galois à partir des données initiales permet de construire une structure sur les données et de montrer leurs relations en regroupant, sous forme de classes recouvrantes, les données présentant des caractéristiques communes. Le treillis nous fournit ainsi un niveau d'abstraction au-dessus des données brutes. Ce niveau est appelé *niveau conceptuel*. A ce niveau, plusieurs treillis peuvent être construits sur différents ensembles de données. Chaque treillis représente un *contexte conceptuel* et constitue un espace de recherche pour l'utilisateur. Nous distinguons le *contexte conceptuel global*, défini par le treillis, et le *contexte conceptuel instantané*, qui dépend du treillis mais aussi de la requête de l'utilisateur ou de sa navigation. Ces deux notions sont présentées dans (Le Grand et al., 2006) et définies formellement dans (Polaillon et al., 2007). Les ensembles de données sur lesquels les différents treillis sont calculés peuvent posséder des intersections non vides. Dans ce cas, une même donnée pourra appartenir à plusieurs contextes conceptuels globaux permettant ainsi une navigation entre différents treillis. Le *niveau sémantique* est construit au-dessus du niveau conceptuel. Il est peuplé d'ontologies dont les concepts sont reliés à des concepts des treillis du niveau conceptuel. L'objectif du niveau sémantique est double : il permet à l'utilisateur de naviguer vers des concepts dont la généralisation n'existe pas dans

les contextes conceptuels globaux ainsi que de passer d'un contexte global à un autre lorsqu'ils ne possèdent pas d'intersection.

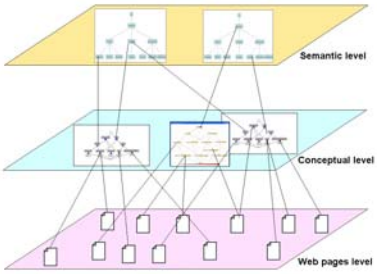

FIG.1 - Trois niveaux de navigation

2.2 Mise en œuvre de l'architecture

La mise en œuvre de l'architecture consiste à sélectionner les ensembles de données à explorer, choisir les ontologies du niveau sémantique, construire les treillis de Galois sur les ensembles de données sélectionnés puis à labelliser les concepts des treillis avec un vocabulaire contrôlé issu des ontologies choisies au niveau sémantique. C'est grâce à cette labellisation que sont construits les liens entre les contextes globaux et les ontologies. Cette mise en œuvre est effectuée hors ligne du fait des temps de calcul engendrés par la construction des treillis et par la labellisation.

Une fois cette étape terminée, l'utilisateur peut explorer les ensembles de données au travers des concepts des treillis du niveau conceptuel ou du niveau sémantique. A un instant *t*, l'utilisateur se trouve sur un concept du niveau conceptuel ou sémantique. A partir du concept courant où se il trouve, l'utilisateur peut naviguer vers un autre concept plus général ou plus spécialisé que le concept courant.

2.3 Outils conceptuels et sémantiques

Dans cette section, nous nous positionnons et rappelons rapidement les outils nous permettant l'utilisation conjointe de treillis de Galois et d'ontologies.

2.3.1 Treillis de Galois

Les treillis de Galois (Godin et al 1993) sont des structures mathématiques permettant l'extraction et la représentation des connaissances. Ils trouvent un usage fondamental en fouille de données.

Soit un contexte formel $K = (O, A, R)$ où O représente l'ensemble des objets, A l'ensemble des propriétés et R une relation binaire entre O et A. La relation R donne lieu à la définition de deux correspondances :

$$Y = f(X) \text{ avec } f(X) = \{a \in A \mid \forall o \in O, oRa\} \; \forall X \subseteq O$$

$$X = g(Y) \text{ avec } g(Y) = \{o \in O \mid \forall a \in Y, oRa\} \; \forall Y \subseteq A$$

Le couple de fonctions f et g est une connexion de Galois entre $P(O)$ et $P(A)$. Un concept (X, Y) est défini par son extension X et son intension Y. L'obtention de tous les concepts

d'une relation binaire par fermeture de la connexion de Galois donne un treillis de Galois (concepts) dont les concepts sont partiellement ordonnés par inclusion ensembliste.

Les treillis de Galois regroupent les données sous forme de concept en fonction de leurs caractéristiques communes et permettent d'exprimer de manière explicite toutes les relations entre les données.

2.3.2 Treillis de Galois et ontologies

Pour améliorer la recherche d'information, nous utilisons les treillis conjointement à des informations sémantiques « extérieures », qui peuvent prendre la forme de taxonomies, de thesaurus ou d'ontologies, selon leur expressivité (Aufaure et al 2006).

Plusieurs travaux se sont intéressés à l'extraction des données et à la représentation des informations sous la forme d'un treillis de Galois.

L'enrichissement des treillis par des ontologies a fait l'objet de nombreux travaux dans le cadre de la recherche d'informations permettant une analyse plus efficace des différentes relations existantes entre les données : enrichissement d'ontologies pour l'organisation et la recherche d'information (Delteil et al 2002) ; les ontologies de domaine ont été utilisées pour guider la construction du treillis selon les préférences de l'utilisateur (Safar et al 2004) et pour enrichir l'indexation du treillis par un thesaurus (Priss 2000).

Ce processus d'extraction a été employé pour divers types de données : gestion de publications (Szathmary et Napoli 2004); aéronautique (Zenou et Samuelides 2004) ; biologie (Messai et al 2005), pages Web (Carpineto et Romano 2004), etc.

Par ailleurs, des travaux ont été réalisés pour que l'utilisateur, face à des treillis de grande taille lors d'une navigation, puisse focaliser son attention sur seulement une partie du treillis par décomposition ou par effet de zoom sur des vues abstraites et plus détaillées (Carpineto et Romano 1995) (Godin et al 1993).

2.4 Discussion

L'architecture à trois niveaux que nous avons présentée permet d'apporter une aide pertinente et contextualisée à la recherche d'information par l'utilisateur. Ce dernier se déplace de manière transparente parmi ces niveaux et passe aisément de l'un à l'autre selon ses besoins en termes de recherche d'information. Nous avons également présenté les qualités qui nous ont conduits à utiliser les treillis de Galois dans notre architecture. Néanmoins, la navigation dans un treillis de Galois présente des difficultés. En effet, le nombre de concepts créés augmente avec le nombre d'individus et le nombre de propriétés, ce qui en fait des structures complexes et très difficiles à interpréter au-delà d'une certaine taille. Compte tenu de cette complexité, une aide est nécessaire pour conseiller l'utilisateur dans le choix du concept de départ de sa navigation et du prochain concept à explorer en termes de pertinence par rapport à sa position actuelle. Pour apporter cette aide, nous avons besoin d'une mesure de similarité entre les concepts d'un treillis ; il nous faut pour cela une mesure de similarité adéquate.

Dans la section 3, nous présentons un état de l'art des mesures de similarité définies dans le cadre de l'appariement d'ontologies et dans la section 4 nous proposons l'extension de l'une d'entre elles pour l'adapter aux treillis de Galois

3 Etat de l'art sur les mesures de similarité pour l'appariement d'ontologies

3.1 Approche basée sur les arcs

Parmi les solutions classiques pour les mesures de similarité, on peut trouver les approches basées sur les arcs qui reposent uniquement sur la structure de l'ontologie. Les deux mesures les plus utilisées sont la mesure de Rada (Rada et al 1989) et celle de Wu & Palmer (Wu et Palmer 1994) décrites ci-dessous. Elles se basent sur la distance en termes de nombre d'arcs séparant un concept d'un autre.

(Rada et al 1989) suggèrent que, pour mesurer la distance entre deux concepts ontologiques, notée *dist (c1, c2)*, on se base sur le nombre d'arcs minimum à parcourir pour aller du concept *c1* au concept *c2*. La mesure de similarité est ainsi de la forme :

$Sim_{rada}(c1, c2) = 1 / 1+dist(c1, c2)$

avec : $dist(c1, c2) = minchemin(c1, c2)$

Dans le même ordre d'idée, (Wu et Palmer 1994) définissent la similarité en fonction de la distance qui sépare deux concepts ontologiques dans la hiérarchie et également par leur position par rapport à la racine. On a ainsi :

$Sim_{W\&P}(c1, c2) = 2*depth(c) / depth(c1) + depth(c2)$.

avec : - *depth(c)* le nombre d'arcs qui séparent le plus petit généralisant de *c1* et *c2* de la racine.

- *depth (ci)* le nombre d'arcs qui séparent le concept *ci* de la racine en passant par *c*.

Ces mesures ont l'avantage d'être faciles à implémenter et peuvent donner une idée sur le lien sémantique entre les concepts. Cependant, elles ne prennent pas en compte le contenu du concept lui-même, ce qui peut conduire, dans certains cas, à une marginalisation de l'apport du concept en terme d'information.

3.2 Approche utilisant le contenu informationnel

Les mesures de similarité suivant cette approche sont fondées sur la notion de Contenu Informationnel (CI) qui utilise conjointement l'ontologie et le corpus. Le Contenu Informationnel d'un concept traduit sa pertinence dans le corpus en tenant compte de sa spécificité ou de sa généralité. Pour ce faire, la fréquence des concepts dans le corpus est calculée et elle regroupe la fréquence d'apparition du concept lui-même ainsi que les concepts qu'il subsume (concepts fils). Les deux mesures les plus connues dans cette catégorie sont celles de Resnik (Resnik 1995) et Jiang-Conrath (Jiang et Conrath 1997).

Resnik (Resnik 1995) définit la similarité sémantique entre deux concepts par la quantité d'information qu'ils partagent : elle est égale au contenu informationnel du concept le plus spécifique (plus petit généralisant *ppg*) qui subsume les deux concepts dans l'ontologie. Elle est définie comme suit :

$Sim(c1, c2) = CI(ppg(c1, c2))$

avec : $CI = -log(P(c))$

où : *P(c)* est la probabilité de retrouver une instance du concept. Ces probabilités sont calculées par la fréquence de *c* sur le nombre total des concepts.

La mesure de (Jiang et Conrath 1997) prend en compte à la fois le contenu informationnel du *ppg* et celui des concepts concernés. Par conséquent, elle peut pallier les limites de la mesure de Resnik et est définie de la manière suivante :

Sim (c1, c2) = 1/distance (c1, c2)

avec : *distance (c1, c2) = CI (c1) + CI (c2) – (2.CI (ppg (c1, c2)))*

D'autres mesures de similarité ont été dérivées de celles présentées ci-dessus dans le cadre d'études assez récentes. Parmi elles, celle de (Zargayouna et Salotti 2004) étend la mesure de Wu et Palmer en ajoutant un calcul de spécificité ; (Blanchard et al 2006) ont défini une nouvelle mesure de similarité qui tend à exploiter au maximum une taxonomie sans considérer le corpus et ce en réutilisant le principe de la quantité d'information.

4 Mesure de similarité adaptée aux treillis de Galois pour l'aide à la navigation

Dans cette section, nous proposons une mesure de similarité permettant d'exploiter à la fois les informations sémantiques et topologiques fournies par l'ontologie et le treillis. Nous expliquons également comment nous mettons cette métrique en œuvre pour faciliter la navigation dans un treillis de Galois.

4.1 Mesure de similarité adaptée aux treillis de Galois

Nous proposons une mesure de similarité entre des concepts d'un treillis de Galois, construit ici pour la classification de pages Web. Cette mesure est une extension d'une mesure de similarité entre les concepts d'une ontologie, que nous avons adaptée à nos besoins, afin de permettre une meilleure étude du treillis. Cette mesure tient compte à la fois de la sémantique et de la topologie, à travers le voisinage d'un concept ainsi que sa profondeur.

L'objectif de cette mesure de similarité est de permettre une meilleure navigation dans le treillis et une restriction du champ des concepts visités. En particulier, nous cherchons à quantifier l'apport d'information des éléments de l'intention d'un concept par rapport à tout le corpus. Nous nous intéressons tout particulièrement à la fréquence des éléments communs à deux concepts, afin de pouvoir avoir une idée sur le contexte partagé.

Chaque concept du treillis de Galois est constitué de deux sous-ensembles (extension-intension). La fréquence des termes de l'intention dans les pages Web constituant le corpus est calculée pour mesurer l'Information Moyenne (IM) de chaque concept, qui intervient, par la suite, dans le calcul de la mesure de similarité entre les concepts. L'IM est dérivée du Contenu Informationnel (CI) défini dans la section 3.2. L'intérêt de l'IM est d'évaluer le poids des termes dans les pages des sites web.

Notre mesure de similarité tient également compte de la topologie du treillis de Galois en faisant intervenir la profondeur de chaque concept *c* (notée *depth(c)* et correspondant au nombre d'arcs qui séparent *c* du concept le plus spécifique du treillis) ainsi que l'information moyenne du plus petit généralisant des deux concepts dont on mesure la similarité.

On définit l'information moyenne d'un concept *c* constitué d'une extension *E* et d'une intention *I* comme le Contenu Informationnel de l'intention, défini de la manière suivante :

$$IM(c) = -\log(P(I))$$

où *P(I)* est la probabilité de retrouver les termes de l'intention (i.e. les termes fréquents) simultanément dans le corpus (i.e. les pages Web). Cette probabilité correspond au rapport entre le nombre de pages Web possédant les termes de I et le nombre total de pages Web du corpus.

Nous nous intéressons ici au calcul de la similarité entre deux concepts du treillis de Galois. Si l'on considère deux concepts *c1* et *c2* du treillis de Galois, on note *ppg (c1,c2)* le plus petit généralisant de *c1* et de *c2*, qui est le « plus proche ancêtre » commun à *c1* et *c2*.

La mesure de similarité entre *c1* et *c2* est alors définie de la manière suivante :

$$Sim(c1,c2) = \frac{1}{IM(c1).depth(c1) + IM(c2).depth(c2) - 2IM(ppg(c1,c2)).depth(ppg(c1,c2))}$$

où *depth(c)* est la profondeur du concept *c*.

Notre mesure est une adaptation de la mesure de Jiang & Conrath (cf. section 3.2). Notre apport consiste à prendre en compte, non seulement l'information moyenne de chaque concept mis en jeu, mais aussi sa profondeur afin de préserver la particularité des treillis : plus on descend au niveau de la hiérarchie et plus on se spécialise. Un autre apport de cette mesure est de tenir compte de la structure du treillis puisqu'elle fait intervenir l'information moyenne du plus proche ancêtre commun aux deux concepts dont on mesure la similarité.

La mise au point d'une telle mesure de similarité au niveau du treillis de Galois permet de retrouver le voisinage du concept requête favorisant la redirection de l'utilisateur au cours du processus de recherche de l'information pertinente, comme nous l'illustrons dans la section 5.

4.2 Méthodologie

La mesure de similarité présentée dans ce papier permet de faciliter la navigation de l'utilisateur à l'intérieur d'un treillis de Galois. Nous distinguons la phase d'initialisation de la navigation et la phase de navigation elle-même.

4.2.1 Initialisation de la navigation

Cette phase a pour objectif de trouver le point d'entrée le plus adapté dans le treillis, c'est-à-dire le concept qui servira de point de départ à la navigation, que nous appelons le *concept de départ* et notons *CD*. Ce concept est constitué d'une extension E_{CD} et d'une intention I_{CD}. La recherche d'information peut être précise, dans le cas où l'utilisateur sait la formuler sous forme de requête, ou non ; nous envisageons ces deux scénarios dans cet article.

A partir d'une requête de l'utilisateur

Dans le cas d'une recherche d'information précise, nous proposons de formuler la requête par R termes de l'ontologie, afin d'utiliser un vocabulaire contrôlé et éviter ainsi les ambiguïtés. Par exemple, si le treillis est construit à partir d'un ensemble de pages Web, les requêtes des utilisateurs pourront être formulées sous la forme d'un ensemble de mots-clés appartenant à ce vocabulaire contrôlé. L'ensemble des mots-clés choisis par l'utilisateur constitue l'intention d'un *concept cible* noté *CT*. Au sein du treillis, le concept de départ,*CD*

choisi pour démarrer la navigation sera le concept dont l'intention possèdera le maximum d'éléments contenus dans l'intention du concept *CT*. Cette intention est constituée de tous les termes de la requête et notée : $I_{CT} = \{p_1, p_2, \ldots, p_R\}$. Si plusieurs concepts du treillis possèdent le même nombre de propriétés recherchées, il faut choisir le concept initial parmi ces concepts dits *candidats*. Pour cela, on calcule leur similarité deux à deux et le concept retenu comme concept de départ est celui dont la valeur moyenne de similarité avec les autres concepts candidats est la plus élevée :

On note $c_1, c_2, \ldots, c_G$ les concepts du treillis, G étant le cardinal du treillis. Soient $cc_1, cc_2, \ldots, cc_N$, N concepts candidats. On note $Sim\ (cc_i, cc_j)$ la similarité entre le concept cc_i et le concept cc_j, avec *i* et *j* appartenant à [1-N]. La similarité moyenne entre un concept candidat cc_k et les autres concepts candidats et dite *partielle*, notée Sim_{ck}^{mp} est égale à la moyenne des valeurs de similarité entre cc_k et les autres concepts candidats. Le concept de départ est celui dont la similarité moyenne avec les autres concepts candidats est la plus élevée, soit :

$CD = cc_i$ *tel que* $Sim_{ci}^{mp} = Max\ (\ Sim_{cj}^{mp})$ pour *j* allant de 1 à N,
avec $Sim_{ci}^{mp} =$ *Moyenne* $(\ Sim\ (cc_i, cc_k)\ )$ pour *k* allant de 1 à N (et k≠i).

En sélectionnant le concept de départ comme étant, en moyenne, le plus proche des autres candidats, on choisit en quelque sorte le concept le plus « central » parmi ceux qui correspondent le mieux à la requête de l'utilisateur. Si M concepts répondent au critère de similarité moyenne maximale, le concept de départ retenu est celui dont la similarité moyenne avec *l'ensemble* des concepts du treillis (y compris les concepts non candidats) est la plus élevée. Pour le concept candidat cc_k, cette similarité moyenne dite *totale* est notée Sim_{ck}^{mt} et correspond à la moyenne des valeurs de similarité entre cc_k et *tous* les autres concepts du treillis. Si l'on note cc_1', cc_2', …, cc_M' les M concepts candidats résiduels, on a :

$CD = cc_i$' tel que $Sim_{ci}{}'^{mt} = Max(\ S_{cj}{}'^{mt})$ pour *j* allant de 1 à M,
avec $Sim_{ci}{}'^{mt} =$ *Moyenne* $(\ Sim\ (\ cc_i', c_k\ )\ )$ pour *k* allant de 1 à G, G étant le cardinal du treillis et k≠i.

Sans requête de l'utilisateur

Si l'utilisateur n'a pas formulé de requête et souhaite simplement naviguer pour explorer un ensemble de pages Web, le concept de départ est celui dont la similarité moyenne totale avec tous les autres objets du treillis est la plus élevée, soit :

$CD=c_i$ tel que $Sim_i^{mt} = Max\ (\ Sim_j^{mt})$ pour *j* allant de 1 à G, G étant le cardinal du treillis,
avec $Sim_i^{mt} =$ *Moyenne* $(\ Sim\ (c_i, c_k)\ )$ pour *k* allant de 1 à G et k≠i.

En sélectionnant le concept de départ comme étant, en moyenne, le plus proche des autres concepts du treillis, on choisit en quelque sorte le concept le plus « central », ce qui semble être un point de départ pertinent et représentatif du treillis.

4.2.2 Navigation

Lorsque l'utilisateur se trouve dans cette phase, il est déjà positionné sur l'un des concepts du treillis : le concept courant qu'il est en train de visiter, noté *CP*. Quand

l'utilisateur désire quitter *CP* pour explorer d'autres concepts, l'objectif est alors de lui proposer le concept le plus proche –sémantiquement et conceptuellement parlant– du concept *CP*. Le concept le plus similaire au concept *CP* semble, en effet, être le choix le plus pertinent pour la poursuite de sa navigation. La similarité entre chaque paire de concepts du treillis de Galois peut être pré-calculée, et il suffit alors d'indiquer à l'utilisateur la valeur de la similarité entre son concept courant et tous les autres concepts du treillis, par exemple par ordre décroissant. En phase de navigation, l'unicité du concept proposé n'est pas une nécessité comme dans la phase d'initialisation et, sous réserve de contraintes ergonomiques, plusieurs concepts candidats peuvent être proposés l'utilisateur.

5 Expérimentation

Dans cette section, nous illustrons la méthodologie présentée précédemment à partir d'un corpus constitué de pages Web relatives au domaine du tourisme.

La première étape consiste à construire les objets et les propriétés qui serviront d'entrée au calcul du treillis de Galois. Dans notre exemple, chaque page Web représente un objet et ses propriétés correspondent aux termes les plus fréquents qu'elle contient. L'extraction des termes les plus fréquents est effectuée à partir d'un thesaurus sur le domaine du tourisme.

Les objets et propriétés ainsi extraits sont rassemblés dans une base de données servant à la construction d'un treillis de Galois. Nous avons implémenté un algorithme incrémental reposant sur celui de (Godin et al 1991) pour construire un treillis où chaque concept est décrit par son extension (ensemble de pages Web) et son intention (termes communs aux pages Web de l'extension). Le treillis obtenu est illustré sur la figure 2.

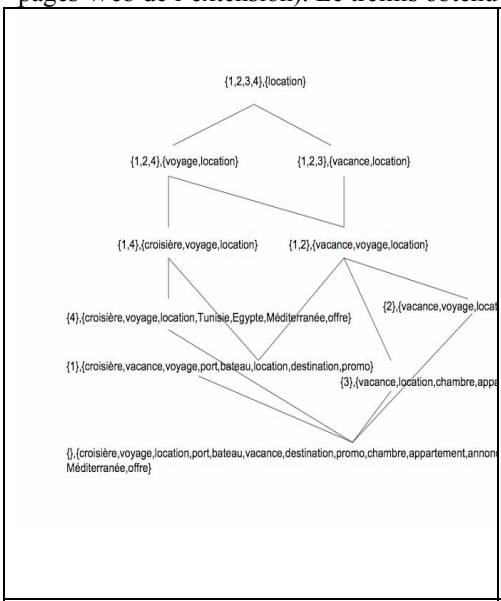

FIG.2 - Treillis de Galois construit à partir de pages Web

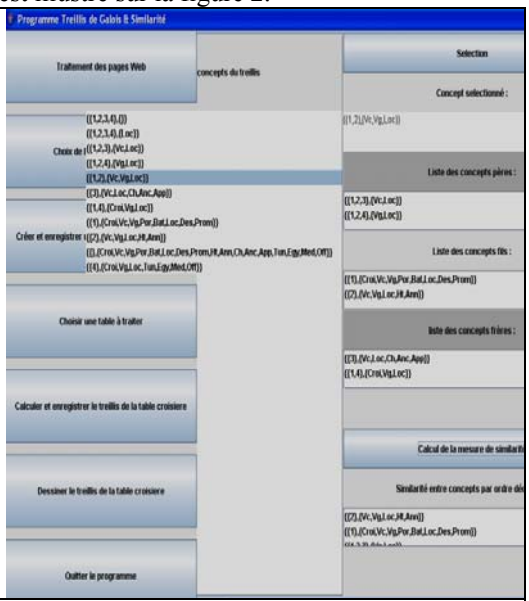

FIG.3 - Calcul de la mesure de similarité à partir d'un concept courant

Supposons que l'utilisateur formule une requête avec les mots-clés suivants : *vacance*, *voyage* et *location*. L'intention du concept cible est donc I_{CT} = {vacance, voyage, location}. Dans ce cas, l'un des concepts du treillis correspond parfaitement : il s'agit du concept *({1, 2} ; {vacance, voyage, location})*, qui va donc être le point de départ de la navigation.

L'utilisateur, s'il souhaite naviguer à partir de ce concept, peut choisir de suivre un arc du treillis pour aller vers un concept plus spécifique ou plus général. Dans cet exemple, il a le choix entre deux pères : *({1, 2, 4), {voyage, location})* et *({1, 2, 3}, {vacance, location})* et deux fils : *({2), {vacance, voyage, location, Hôtel, annuaire})* et *({1}, {croisière, vacance, voyage, port, bateau, location, destination, promo})*. Plusieurs questions se posent à l'utilisateur : comment choisir entre plusieurs pères sans devoir faire un choix a priori sur les propriétés retenues ? Comment choisir entre plusieurs fils sans devoir faire un choix a priori sur les pages retenues ? Par ailleurs, n'est-il pas possible que le concept le plus pertinent pour poursuivre sa navigation ne soit ni un père, ni un fils du concept courant, c'est-à-dire un concept qui n'est pas relié à celui-ci par un arc du treillis (par exemple un frère tel que le concept *({1, 4), {croisière, voyage, location}))* ?

Afin de répondre à toutes ces questions, on regarde la valeur de la mesure de similarité entre le concept courant et tous les autres concepts du treillis, afin de choisir le concept le plus pertinent pour poursuivre la navigation à partir du concept que l'utilisateur est en train de visiter. Toutes ces valeurs sont présentées à l'utilisateur par ordre décroissant comme le montre la figure 3. En l'occurrence, le concept le plus similaire est le concept *({2), {vacance, voyage, location, hôtel, annuaire})*. Il est à noter que toutes les valeurs de similarité entre chaque paire de concepts peuvent être pré-calculées afin d'optimiser le temps de réponse.

6 Conclusion et perspectives

L'extraction d'informations pour en faciliter la recherche peut être réalisée de différentes manières : par des techniques de clustering numériques, basées sur la fréquence d'apparition des termes dans un document, ou par des techniques de clustering conceptuel, permettant d'effectuer des regroupements d'objets partageant les mêmes propriétés. L'avantage de ces dernières est de permettre une structuration des données, et d'offrir des mécanismes de généralisation/spécialisation bien adaptés à l'utilisateur final. Un niveau sémantique, ontologie ou thesaurus, a été proposé au dessus du niveau conceptuel dont les avantages ont été détaillés en section 2.

La problématique abordée dans cet article est de trouver un moyen d'indiquer à l'utilisateur les concepts les plus pertinents par rapport à sa requête, ou de pouvoir lui proposer un point d'entrée dans la structure conceptuelle (treillis de Galois) qui peut être de grande taille. Nous avons donc proposé une mesure de similarité tenant compte à la fois de la sémantique et de la topologie du treillis (position du concept dans le treillis, prise en compte du voisinage). Cette mesure permet donc d'ordonner l'ensemble des concepts se trouvant dans le voisinage d'un concept donné dans le treillis, et ainsi de guider la navigation de l'utilisateur.

Cette mesure a été expérimentée sur des treillis construits à partir de pages Web, les propriétés communes étant des termes extraits de ces pages. Les résultats ont été comparés avec d'autres mesures comme celle de Wu&Palmer, décrite en section 3. Les résultats obtenus montrent un meilleur ordonnancement du voisinage avec notre mesure de similarité.

Les perspectives de ce travail consistent dans un premier temps à développer une interface permettant à des utilisateurs de valider cette approche de manière expérimentale. Nous travaillons également sur la visualisation spatiale des corpus de documents à la base de la construction des treillis, ainsi que sur la manière de mettre en correspondance ces régions spatiales avec les concepts du treillis, pour une meilleure interprétation du voisinage des concepts. Dans la perspective de traiter de gros volumes de données, on pourra aussi s'intéresser à la manipulation de gros volumes de données avec les treillis de Galois grâce à l'utilisation d'algorithmes de construction de treillis récents tel que proposé par (Diday et Emilion, 2003).

Références

Aufaure, M.-A., Le Grand, B., Soto, M. and Bennacer, N. (2006) Metadata- And Ontology-Based Semantic Web Mining, In *Web Semantics And Ontology*, D. Taniar & J. Wenny Rahayu Eds., Idea Group Publishing, pp. 259-296.

Blanchard, E., Harzallah, M., Kuntz, P. and Briand, H. (2006) Une nouvelle mesure sémantique pour le calcul de la similarité entre deux concepts d'une même ontologie. Revue nationale des nouvelles technologies de l'information.

Carpineto, C., Romano, G. (1995) Automatic construction of navigable concept networks characterizing text databases. Topics in Artificial Intelligence.

Carpineto, C., Romano, G. (2004) Exploiting the Potential of Concept Lattices for Information Retrieval with CREDO. Journal of Universal Computer Science, vol. 10, n. 8, pp. 985-1013.

Delteil, A., Faron, C., Dieng, R. (2002) Building Concept Lattices by Learning Concepts from RDF Graphs Annotating Web Documents. Proceedings of the 10th International Conference on Conceptual Structures: Integration and Interfaces.

Diday, E., Emilion, R. (2003) Maximal and Stochastic Galois Lattices. Discrete Applied Mathematics 127(2): 271-284

Godin, R., Missaoui, R., Alain, A. (1993) Experimental comparison of navigation in a Galois lattice with conventional information retrieval methods. International Journal of Man-Machine Studies, 38, 747-767.

Godin, R., Missaoui, R. and Alaoui, H. (1991) Learning algorithms using structure. In IEEE Int. Conf. on Tools for Artificial Intelligence.

Jiang, J. et Conrath, D. (1997) Semantic similarity based on corpus statistics and lexical taxonomy. In. Proceedings on International Conference on Research in Computational Linguistics, Taiwan.

Le Grand, B., Aufaure, M.-A., Soto, M. (2006) Semantic and Conceptual Context-Aware Information Retrieval, the IEEE/ACM International Conference on Signal-Image Technology & Internet-Based Systems (SITIS'2006), pp. 322-332, Hammamet, Tunisie, 17-22 décembre 2006.

Messai, N., Devignes, MD., Napoli, A., Tabbone, M-S. (2005) Treillis de concepts et ontologies pour l'interrogation d'un annuaire de sources de données biologiques (BioRegistry). 18 ème Congrès INFORSID 2005.

Polaillon, G., Aufaure, M.-A., Le Grand, B., Soto, M**.** (2007) FCA for contextual semantic navigation and information retrieval in heterogeneous information systems. Workshop on Advances in Conceptual Knowledge Engineering, in conjunction with DEXA 2007, Regensburg, Allemagne, pp. 534-539.

Priss, U.(2000) Lattice-based information retrieval. Knowledge Organization, 27(3):132–142.

Rada, R., Mili, H., Bicknel, E., Blettner, M. (1989) Development and application of a metric on semantic nets. IEEE Transaction on Systems, Man, and Cybernetics, 19(1):17--30.

Resnik, P. (1995) Using information content to evaluate semantic similarity in a taxonomy. In Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal.

Safar, B., Kefi, H., Reynaud, C. (2004) OntoRefiner, a user query refinement interface usable for Semantic Web Portals, Application of Semantic Web Technologies to Web Communities (ECAI'2004) August 23rd, Spain, 16th European Conference on Artificial Intelligence, August 22-27, 2004, Valencia (Spain), p65-p79.

Szathmary, L., Napoli, A. (2004) Knowledge Organization and Information Retrieval with Galois Lattices. EKAW 2004: 511-512.

Wu, Z. et Palmer, M. (1994) Verb Semantics and Lexical Selection, Proceedings of the 32nd Annual Meetings of the Associations for Computational Linguistics, pages 133-138.

Zargayouna, H., Salotti, S. (2004) Mesure de similarité sémantique pour l'indexation de documents semi-structurés dans 12ème Atelier de Raisonnement à Partir de Cas.

Zenou, E., Samuelides, M. (2004) Utilisation des treillis de Galois pour la caractérisation d'ensembles d'images. 14ème Congrès Francophone AFRIF-AFIA de Reconnaissance des Formes & Intelligence Artificielle (RFIA'2004), Toulouse (France), Vol.1, pp.395-404.

Summary

Information retrieval and navigation in Web pages are complex tasks because of the growing volume of data and their lack of structure. Conceptually formalizing a *context* associated to an ontology allows to enhance this process. We define a *conceptual context* as the association of ontologies and Galois lattices built from Web pages. Our goal is to propose a similarity measure allowing users to navigate more easily in lattices. In order to enhance navigation, the most relevant concepts with regard to a specific request should be shown to the user, or the best starting point for a '"free" navigation. Our contribution relies on the fact that we consider a semantic link between the lattice's concepts, based on an extension of the similarity measures used for ontologies.