Extraction et exploitation de données temporelles pour un portail d'e-tourisme

Jérôme Fortin, Olivier Carloni, Michel Leclère, Stéphanie Weiser

Extraction et exploitation de données temporelles pour un portail d'e-tourisme

Jérôme Fortin*, Olivier Carloni **, Michel Leclère*, Stéphanie Weiser***

*LIRMM, 161 rue Ada 34392 Montpellier Cedex 5 - France
{fortin,leclere}@lirmm.fr
**Mondeca, 3, cité Nollez 75018 Paris France
olivier.carloni@mondeca.com
***MoDyCo, 200, avenue de la République 92001 Nanterre Cedex - France
sweiser@u-paris10.fr

Résumé. Dans le cadre d'un portail sémantique d'e-tourisme, des données sont collectées depuis des pages web par un processus d'acquisition automatique afin de renseigner une base de connaissances touristique. Une partie cruciale des connaissances à acquérir concerne des informations temporelles sur les dates et horaires d'ouvertures des ressources touristiques. Ces informations d'ouverture sont souvent données de manière incomplète, vague, ambiguë et leur interprétation fait appel à des connaissances implicites du domaine. Dans cet article, nous présentons le processus d'acquisition mis en œuvre en expliquant les problèmes posés par ces informations temporelles. Nous proposons alors d'utiliser une modélisation possibiliste de ces connaissances afin de tenir compte, d'une part, de l'imprécision des connaissances extraites, et d'autre part, de permettre souplesse et flexibilité dans leur exploitation. Nous discutons finalement des raisonnements envisagés, en montrant comment exploiter les informations temporelles ainsi modélisées.

1 Introduction

Le projet Eiffel[1] a pour ambition le développement de portails sémantiques d'e-tourisme permettant à un territoire donné (i.e. un acteur institutionnel d'un bassin touristique) de valoriser son offre touristique en proposant un accès unifié et recomposé aux diverses offres des prestataires de tourisme dépendant de ce territoire. La solution développée se décompose en deux phases (cf. Noël et al. (2008)) : l'une dédiée à l'acquisition de l'offre touristique du territoire dont l'objectif est l'élaboration d'une base de connaissances tourisme/territoire, l'autre dédiée à la présentation de cette offre via un portail sémantique exploitant cette base de connaissances.

La base de connaissances s'appuie sur une ontologie tourisme/territoire adaptée au territoire visé. La phase d'acquisition dispose de diverses modalités de saisie pour "peupler" cette ontologie : saisie manuelle par les acteurs du territoire (par exemple les offices de tourisme), importation automatique de catalogues de grands prestataires touristiques, mais aussi acquisition automatique par exploration du web et construction automatique d'annotations. C'est à cette dernière modalité que nous nous intéressons dans cet article. Elle présente l'intérêt de capter les nombreuses offres temporaires en particulier toutes celles touchant à l'événementiel : concerts, expositions, festivals, foires... En effet, ces offres ne sont en général pas décrites dans les catalogues "classiques" et/ou sont très mal mises à jour. Cette modalité d'acquisition automatique est réalisée en 3 étapes : (1) le web est exploré à la recherche de pages présentant des offres pertinentes pour le territoire visé, (2) une phase d'extraction automatique des caractéristiques de l'offre est alors effectuée sur chaque page identifiée et permet d'obtenir une première annotation "brute" de l'offre touristique, (3) ces annotations sont ensuite analysées, enrichies et validées avant d'être intégrées à la base de connaissances.

Deux caractéristiques sont essentielles lors de la description d'une offre touristique : sa localisation et sa période de validité. Dans cet article, nous nous focalisons sur l'acquisition des caractéristiques temporelles de l'offre. Une première analyse des données extraites des pages web a montré que les expressions linguistiques utilisées pour décrire ces caractéristiques temporelles sont variées, incomplètes, vagues et parfois ambiguës. De plus, l'interprétation de ces expressions nécessite la mobilisation de connaissances implicites liées au domaine (comme par exemple, la notion de haute-saison, les dates des vacances scolaires, les horaires "classiques" d'ouverture de certains types de ressources (les discothèques sont ouvertes la nuit, les musées le jour...)). L'objectif de ce premier travail est de permettre de déterminer automatiquement les périodes d'ouverture d'une offre touristique à partir d'une description textuelle récupérée sur une page web.

Après avoir expliqué les principes de fonctionnement de la phase d'analyse, nous proposons d'utiliser un modèle possibiliste pour les données temporelles afin de prendre en compte l'imprécision des données. Nous montrons alors comment une telle modélisation peut être mise en œuvre dans le cadre du projet Eiffel.

[1]Ce travail a été financé par le projet RNTL-ANR Eiffel (cf. http ://www.projet-eiffel.org).

Dans la Section 2, nous détaillons les caractéristiques des données à analyser et proposons un modèle d'annotation pour les connaissances brutes extraites des expressions linguistiques. En Section 3, nous expliquons les principes de fonctionnement de la phase de construction automatique de ces annotations. La Section 4 présente les objectifs de la phase d'analyse des annotations en exhibant les résultats attendus et en introduisant la méthodologie utilisée. Enfin, la Section 5 propose une modélisation possibiliste des données et discute des avantages de cette approche prometteuse.

2 Caractéristiques des données à traiter

Les expressions temporelles que l'on souhaite repérer et annoter ont une visée informative et pratique. Il ne s'agit pas de dates historiques ou d'expressions descriptives du type *la nuit d'avant* mais d'informations pratiques dans le domaine du tourisme. Il peut ainsi s'agir d'horaires d'ouverture, de dates, de périodes, etc. On peut classer ces expressions en deux catégories principales (cf. Weiser et al. (2008)) : les informations temporelles qui concernent un événement particulier et les informations temporelles répétitives. La première comprend des dates (*concert le 1er octobre*), des périodes (*festival de mai à Juin*), des heures (*le concert commence à 8h*). La seconde comprend des horaires (*le musée ouvre à 10h*), des périodes (*le restaurant est ouvert du lundi au samedi*) et des exceptions (*le camping est ouvert toute l'année sauf en janvier*). Des exemples d'une complexité plus grande peuvent également prendre place dans cette classification comme *de mai à Juin, ouvert tous les jours sauf le mardi*.

Les expressions temporelles touristiques peuvent être génériques ou propres à un type d'objet en particulier. Par exemple une date comme *le 31 octobre 2008* peut aussi bien convenir à un concert, qu'à une représentation de théâtre ou à l'ouverture d'une patinoire. En revanche, une expression comme *ouvert midi et soir sauf le lundi* peut difficilement s'appliquer à autre chose qu'à un restaurant. Ces exemples montrent que la complexité des expressions temporelles varie énormément : certaines expressions sont très simples, d'autres peuvent devenir complexes, jusqu'au point d'être floues ou ambiguës. Par exemple dans *vendredi et samedi soir*, doit-on comprendre qu'il s'agit des deux soirées ou de la journée de vendredi et de la soirée de samedi ? Le contexte est nécessaire pour lever ce type d'ambiguïtés. Pour cet exemple, s'il s'agit d'un concert, la première interprétation sera probablement la bonne.

Certaines expressions présentent donc des difficultés d'interprétation : ambiguïtés virtuelles ou même réelles[2], dates imprécises, etc. Dans d'autres cas, si les expressions sont naturellement interprétées par un internaute, leur interprétation a recours à des connaissances du monde ou nécessite des inférences (par exemple, lorsque l'on a des horaires de fermeture et que l'on peut naturellement en déduire des horaires d'ouverture). Voici un inventaire plus précis de ces difficultés :

- *Fermé le mardi.* Pour une telle expression, le but est de déduire les jours d'ouverture (donc à priori lundi, mercredi, jeudi, vendredi, samedi et dimanche). Si on dispose d'informations complémentaires concernant le type d'objet, on peut éventuellement ajouter des parties de journées : si c'est une boîte de nuit, considérer pour chaque jour d'ouverture qu'il s'agit de la nuit ou s'il s'agit d'un restaurant, déduire des parties de journées comme midi et soir.
- *Visites de Juin à septembre les après-midi, sauf lundi et mardi.* Cette expression est ambiguë : il n'est pas possible de savoir si le lundi et le mardi la ressource est fermée toute la journée ou si, au contraire elle est ouverte le matin. De plus, elle est également floue au niveau des dates : il peut s'agir du 1er Juin ou du premier week-end de Juin ou autre. Et pour septembre, il peut s'agir du début ou de la fin du mois.
- *Fermé durant les vacances scolaires de février. Fermeture hebdomadaire non déterminée.* Avec cette expression, on ne peut déduire aucun jour d'ouverture ou de fermeture. Pour interpréter les périodes de fermeture (et ensuite en déduire les périodes d'ouverture par complémentarité), il faut faire appel à une connaissance du monde : les dates de vacances scolaires. Il est possible d'utiliser pour chaque année un calendrier comprenant les vacances.

Ces différents exemples donnent donc un petit panorama des calculs qui doivent être effectués sur les expressions effectivement repérées dans les pages web de ressources touristiques.

3 Construction de l'annotation

En amont de la phase de construction de l'annotation, une première étape de "crawling" (effectuée par l'un de nos partenaires[3]) se charge de collecter des pages web à caractère touristique (sites d'hôtels, de restaurants, d'événements ponctuels, spectacles, concerts, etc.). Ces pages, au format HTML, sont alors transformées en documents XML, format plus adéquat pour un traitement automatique, et "nettoyées". C'est-à-dire que seules les informations utiles à notre analyse ont été conservées. De nombreuses balises ont donc été supprimées. Il faut noter que, une fois à l'entrée de l'étape d'annotation automatique, chaque page est indépendante : il se peut que plusieurs pages soient issues du même site mais,

[2] On parle d'ambiguïté réelle quand l'expression présente plusieurs interprétations possibles, aussi bien pour un locuteur natif que pour une machine. Une expression virtuellement ambiguë n'est ambiguë que pour une machine.

[3] La société Antidot, partenaire du projet, développe AFS (http ://www.antidot.com) un moteur de recherche adapté à cette tâche.

une fois l'aspiration effectuée, elles ne sont plus liées. Ces pages web sont ensuite analysées de façon à y détecter et à annoter automatiquement les informations utiles à la base de connaissances. Nous nous focalisons ici sur le repérage et l'annotation des informations d'ouverture et de fermeture.

L'objectif de cette première étape est de construire une annotation contenant uniquement les informations explicites contenues dans la page. C'est-à-dire que si la page contient l'expression *"ouvert du lundi au jeudi"*, c'est celle-ci qui sera annotée sans interprétation du type *"ouvert le lundi, le mardi, le mercredi, le jeudi"*. De la même manière, si l'on a *"fermé le mardi"*, cela ne sera pas converti en jours d'ouverture, mais cela sera annoté comme un jour de fermeture. De plus, certaines informations n'ont pas besoin d'être annotées. En effet, seules celles qui peuvent être utiles à l'utilisateur et qui peuvent prendre place dans la base de connaissances doivent être prises en compte. Pour cela, nous nous basons sur l'ontologie tourisme/territoire élaborée pour les besoins du projet.

Comme nous l'avons vu dans la partie précédente, les informations temporelles à annoter sont des périodes d'ouverture ou de fermeture qui peuvent être constituées : d'une date de début et d'une date de fin, d'une date seule, d'une heure de début et d'une heure de fin, de jours. Afin de faciliter ensuite l'intégration des données annotées à la base de connaissances, une DTD[4] définissant le format d'annotation a été établie (une première version est décrite dans Weiser (2008)). Les expressions temporelles peuvent être annotées avec les balises : `période-ouverture`, `période-fermeture`, `exception` et `incertitude`. La balise `exception` permet d'annoter la chaîne textuelle décrivant une exception (comme *sauf le mardi*) et ainsi de garder l'information telle quelle afin de la fournir textuellement à l'utilisateur. La balise `incertitude` permet d'indiquer que le résultat n'est pas fiable : il est flou ou comprend une ambiguïté.

En ce qui concerne les balises `période-ouverture` et `période-fermeture`, elles permettent de définir plus précisément l'expression repérée et peuvent inclure les balises `date`, `date-début`, `date-fin`, `jour`, `heure-début`, `heure-fin` et `partie-de-journée`. La balise `date` sert à annoter les dates seules ; dans la base de connaissances, on considèrera que la date de fin est alors la même que la date de début. La balise `jour` permet d'annoter les jours de la semaine tandis que `heure-début` et `heure-fin` annotent les heures et que `partie-de-journée` annote les informations du types *matin* et *après-midi*. À terme, dans la base de connaissances, toutes les informations d'horaires seront converties en parties de journées.

Pour repérer et annoter les informations temporelles, notre système est basé sur une approche symbolique, reposant sur des patrons linguistiques. Pour chaque type d'information à repérer (informations temporelles, informations spatiales, informations sur le type de la ressource), un module de transducteurs a été développé à l'aide de l'outil Unitex[5]. Cet outil permet de traiter des corpus en utilisant des dictionnaires (basés sur les tables du LADL[6]) ; et ce au niveau du lexique, de la syntaxe ou de la morphologie. Il permet de repérer et de baliser des structures correspondant à des expressions régulières, représentées par des graphes à états finis. A titre indicatif, le module de repérage et d'annotation temporel regroupe 24 graphes comprenant 88 marqueurs de surface (comme *"ouvert"*, *"mardi"*, etc.) et 22 marqueurs généraux (*"le", "du"*, etc.). La sortie d'Unitex est stockée dans un fichier texte dans lequel des balises d'annotations, au format XML, ont été ajoutées pour marquer les données identifiées. Ainsi, pour l'expression *Ouvert Juillet Août, sauf jours fériés du Mardi au Dimanche*, on obtiendra l'annotation suivante :

```
<Periode-Ouverture>
        <Date-Debut> Juillet </Date-Debut>
        <Date-Fin> Août </Date-Fin>
        <Incertitude/>
        <Exception> sauf jours fériés </Exception>
        <Jour>du mardi au dimanche<Jour>
</Periode-Ouverture>
```

4 Enrichissement de l'annotation

L'étape de construction a permis de recueillir les périodes de fermeture et d'ouverture qui sont explicitement précisées dans la page web de la ressource. L'objectif de cette deuxième étape est de prendre en compte les connaissances implicites afin de les rendre explicites dans les informations intégrées à la base. D'une part, une ressource peut présenter uniquement des périodes de fermeture, uniquement des périodes d'ouverture ou bien combiner ces deux types d'informations. On est par exemple confronté à une information du type *"la discothèque l'Oiseau Noir est fermée les dimanche, lundi, mardi et mercredi"* qui signifie d'une façon naturelle qu'elle est ouverte les autres jours de la semaine. Il est donc nécessaire

[4] Une DTD (Document Type Definition) permet de décrire un modèle de document XML.

[5] Unitex : http ://www-igm.univ-mlv.fr/ unitex

[6] Laboratoire d'Automatique Documentaire et Linguistique - les tables ont été créées au LADL par Maurice Gross et contiennent des unités lexicales classées selon des propriétés syntaxiques et distributionnelles.

d'envisager de calculer des périodes d'ouverture à partir de périodes de fermeture. D'autre part, certaines périodes de fermeture peuvent ne pas être explicitement spécifiées car elles sont intuitivement déductibles du contexte associé à la ressource. Par exemple, si la ressource touristique se situe en France, on en déduit qu'elle est fermée les jours fériés. Dans ce cas, on se base sur les propriétés géographiques de la ressource pour déterminer la période de fermeture. Ou encore, si la ressource est une discothèque, elle sera fermée en journée. Dans ce cas, on s'est basé sur la nature de la ressource. De manière générale pour ces deux cas, on s'appuie sur des connaissances qui relèvent du domaine pour déduire une période de fermeture.

L'enrichissement des annotations est réalisée en trois temps :

1. Dans un premier temps, l'annotation subit un traitement au cours duquel toute période est traduite en une donnée facilement exploitable par le mécanisme décrit dans la suite. Ce traitement permet de *désambiguïser* et de *discrétiser* les informations temporelles. Par exemple, *"début janvier"* sera traduit en *"01/01"*, ou bien *"début de la semaine"* en *"lundi"*. Une période est discrétisée en des moments de la journée pour chaque jour de la période (matin, midi, après-midi, soir, nuit). Dans le cadre d'un portail sémantique tourisme/territoire, il parait inutile d'être plus précis.
2. Dans un second temps, on se base sur les propriétés de la ressource pour déterminer ses périodes de fermeture par défaut. Pour ce faire, un ensemble de règles spécifiant les connaissances implicites de domaine est intégré à l'ontologie. Ces règles associent une période de fermeture à un profil de ressource particulier (par défaut une ressource est présumée ouverte tout le temps). Ainsi on peut exprimer des règles du type : *si une ressource est une discothèque alors on en déduit qu'elle est fermée les matins, les midis et les après-midis de tous les jours de la semaine*, ou encore, *si une ressource est située en France alors elle est fermée les jours fériés (ces jours étant explicitement énumérés)*. Ces règles sont appliquées sur les ressources par un service de raisonnement développé dans le cadre d'Eiffel (ce service est décrit dans Carloni et al. (2007)). Ces règles pouvant générer des contradictions, une fois leur application terminée, le service de raisonnement vérifie la cohérence de l'annotation *enrichie* : si une période de fermeture inférée chevauche une période d'ouverture initialement présente dans l'annotation, alors le contenu inféré de l'annotation n'est pas conservé. En cas de contradiction, les connaissances de la page web sont donc considérées comme prioritaires par rapport à celles qui ont été inférées.
3. Dans un troisième temps, on homogénéise le contenu de la base en ramenant toutes les connaissances à des périodes d'ouverture. En effet, l'exploitation de la base de connaissances portera uniquement sur l'ouverture et non sur les périodes de fermeture. Pour ce faire, on s'appuie sur les périodes de fermeture que présente la ressource : on détermine la période de référence que cette période de fermeture impacte, et on en déduit la période d'ouverture comme étant le complémentaire de la période de fermeture par rapport à cette période de référence. La période de référence est la période d'ouverture de la ressource lorsqu'elle en présente une, sinon on choisit celle qui convient le mieux et qui englobe la période de fermeture parmi une liste de périodes de référence par défaut. Par exemple, si la phase d'acquisition a permis d'extraire *la discothèque l'Oiseau Noir qui est fermée les dimanche, lundi, mardi et mercredi* ; lors de l'application des règles, on a déduit que cette discothèque était aussi fermée le matin, le midi et l'après-midi de tous les jours de la semaine. A l'étape d'homogénéisation, comme la période de fermeture est exprimée à la fois en jours de la semaine et en moments de la journée et qu'aucune période d'ouverture n'est spécifiée, on considère que la période de référence est la semaine découpée en moments de la journée. La période d'ouverture obtenue concerne tous ces moments à l'exclusion de ceux appartenant à la période de fermeture : soit jeudi soir/nuit, vendredi soir/nuit et samedi soir/nuit.

L'approche mise en œuvre présente le désavantage de "gommer" certaines nuances en établissant des choix arbitraires qui mériteraient une analyse plus fine. Par exemple, supposons que le processus d'acquisition extrait un restaurant fermé les lundi/mardi et ouvert le jeudi. L'approche décrite ici détermine qu'il est ouvert du mercredi au dimanche ce qui conduit à éliminer l'ambiguïté existant dans l'énoncé initial sur les mercredi, vendredi, samedi et dimanche. Il pourrait être intéressant de considérer cette ambiguïté comme une information à part entière sur laquelle peut s'appuyer le système d'interrogation. On peut aussi souhaiter introduire plus de nuances dans la définition d'une règle de déduction de périodes de fermeture. Par exemple, dire qu'*il y a de fortes chances pour qu'un restaurant soit fermé le lundi* est sans doute préférable à *un restaurant est toujours fermé le lundi*. Ce type d'ambiguïtés et de degré de certitude dans l'expression d'une connaissance est difficilement exprimable dans l'approche actuelle. La section suivante présente une seconde approche qui permet de mieux répondre à ce genre de besoins.

5 Proposition de modélisation possibiliste d'expressions temporelles

Au lieu de fixer de manière arbitraire les périodes d'ouverture d'une ressource touristique au coût d'une interprétation optimiste ou pessimiste (par exemple en interprétant *Ouvert de fin Janvier à fin Février* par *ouvert dès le 30 janvier* ou *ouvert à partir du 2 février*), on peut opter pour l'utilisation de la théorie des possibilités Dubois et Prade (1988) qui,

d'une part, permet de définir une certaine gradualité dans la définition d'une date mal connue d'ouverture ou fermeture et, d'autre part, permet de modéliser un manque d'information de manière réaliste.

Le principe est d'affecter à une ressource pour chaque date, un degré de possibilité d'ouverture μ_O et un degré de possibilité de fermeture μ_F, dont les valeurs peuvent aller de 0 à 1 selon qu'il est respectivement impossible ou possible que la ressource touristique soit ouverte ou fermée. Avec $0,8$ par exemple, on dit que la connaissance est possible à un degré $0,8$. Pour garantir une consistance dans nos données, il faut que l'ouverture ou la fermeture d'un établissement à une date donnée soit complètement possible (i.e. pour chaque date l'un des degrés est à 1), sinon, les informations sont considérées incohérentes. Ceci correspond à une distribution de possibilité normée, dans laquelle au moins une des valeurs envisagées est complètement possible (ici nous avons deux valeurs envisageables qui sont *ouvert* et *fermé*). Notons au passage qu'en cas de manipulation d'informations incohérentes, il est toujours possible de normaliser les degrés de possibilités avant l'utilisation de ces données (par exemple, en ramenant à 1 le degré le plus grand).

Si on suppose qu'une ressource est fermée un jour, mais qu'on imagine tout de même possible qu'elle soit en réalité ouverte (par exemple un restaurant est supposé fermé le lundi soir, mais on sait qu'il n'est pas impossible qu'un restaurant soit tout de même ouvert), on peut modéliser notre connaissance par les degrés de possibilité $\mu_O = 0.5$ et $\mu_F = 1$.

Il est alors intéressant de définir un degré de certitude pour notre information (appelé plus souvent mesure de nécessité). Dans nos exemples, le degré de certitude concernant l'ouverture d'une ressource est défini par 1 moins la possibilité qu'elle soit fermés. Dans l'exemple précédent, on ne peut donc pas être sûr que le restaurant en question soit ouvert. En revanche on est sûr à un degré 0.5 qu'il est fermé. Voici par exemple pour un établissement et un jour donnés, certaines valeurs que peuvent prendre les deux degrés, et les conclusions que l'on peut en tirer :

- $\mu_O = 1$ et $\mu_F = 0$: on est alors sûr que l'établissement est ouvert ce jour la.
- $\mu_O = 0$ et $\mu_F = 1$: on est alors sûr que l'établissement est fermé ce jour la.
- $\mu_O = 1$ et $\mu_F = 1$: on est en présence d'une incertitude totale, on ne peut rien conclure.

5.1 Exemple de modélisation d'une connaissance vague

Prenons l'exemple de l'expression *Fermeture annuelle de fin Janvier à fin Février*. Plutôt que de déterminer "fin janvier" arbitrairement comme le "31 janvier", on peut modéliser l'ouverture de l'établissement par les distributions de possibilités représentées sur la Figure 1. On définit formellement pour chaque date une distribution de possibilité normée sur l'ensemble des éléments $\{ouverture, fermeture\}$, mais pour des questions de lisibilité nous représentons ces informations sous la forme de deux fonctions $\mu_O(.)$ et $\mu_F(.)$ qui donnent les degrés de possibilité des éléments *ouvert* et *fermé* à une date donnée. Du 25 janvier au 31 janvier, il est de de moins en moins pensable que l'établissement soit ouvert, le degré de possibilité d'ouverture diminue donc progressivement de 1 à 0 entre le 25 et le 31 janvier. Inversement, du 20 au 25 janvier, il est de plus en plus pensable que l'établissement soit fermé.

L'établissement en question est-il ouvert le 28 janvier ? On déduit de la distribution de possibilités d'ouverture qu'il est possible à un degré 0.3 (donc très moyennement possible) que l'établissement soit ouvert à cette date là. En revanche, on peut avancer que l'établissement est fermé avec un degré de certitude de 0.7.

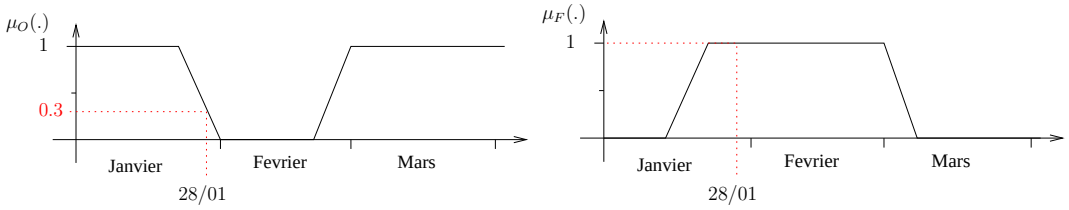

FIG. 1 – *Fermeture annuelle de fin janvier à fin février*

5.2 Exemple de modélisation d'une connaissance ambiguë

Soit l'expression *Ouvert les après-midi sauf lundi et mardi*. Dans cette expression, il est certain que l'établissement est ouvert les après-midis du mercredi au dimanche (degrés de possibilité d'ouverture = 1 et de fermeture = 0). On est sûr qu'il est fermé les lundi après-midi et mardi après-midi (degré de possibilité d'ouverture = 0 et de fermeture = 1), et l'ouverture est incertaine pour le reste du temps puisqu'assujettie à l'interprétation de l'expression. On peut modéliser cette incertitude en affectant un degré de possibilité d'ouverture 0,5 aux matinées (avec un degré de possibilité de fermeture de 1). Ceci nous donne les distributions de possibilités d'ouverture et de fermeture résumées dans la Table 1.

Ainsi à la requête "l'établissement est-t-il ouvert toute la journée du mercredi ?", la réponse serait oui avec un degré de certitude de 0. À la requête "l'établissement est-t-il ouvert à un moment de la journée du mercredi ?", la réponse serait oui avec un degré de certitude de 1. À la requête "l'établissement est-t-il ouvert le lundi ?", la réponse serait oui avec un degré de possibilité de 0.5. Et enfin à la requête "l'établissement est-t-il fermé le lundi ?", la réponse serait oui avec un degré de possibilité de 1 et un degré de certitude de 0.5.

		Lundi	Mardi	Mercredi	Jeudi	Vendredi	Samedi	Dimanche
μ_O	matin	0.5	0.5	0.5	0.5	0.5	0.5	0.5
	soir	0	0	1	1	1	1	1
μ_F	matin	1	1	1	1	1	1	1
	soir	1	1	0	0	0	0	0

TAB. 1 – *Ouvert les après-midi sauf lundi et mardi*

5.3 Exemple de modélisation de connaissances contextuelles

Les connaissances contextuelles peuvent permettre d'obtenir le contour des distributions de possibilités de nos connaissances. Par exemple, *"en général, les musées sont fermés le soir"*, *"les restaurants sont fermés les lundi soir"*.

On peut affecter un degré de possibilité d'ouverture (resp. fermeture) de 1 lorsqu'il est contextuellement supposé qu'un établissement est ouvert, et un degré de possibilité de l'événement contraire fermeture (resp. ouverture) de 0.5. Sans autre source d'information que le contexte, la connaissance est donc incertaine. Pour un restaurant, cela donne les connaissances contextuelles définies dans la Table 2.

S'il est précisé sur le site internet d'un restaurant *ouvert le lundi soir*, il faut alors changer la possibilité d'ouverture du lundi par 1, et la possibilité de fermeture par 0 (on ne pourra affirmer que l'établissement est ouvert le mercredi qu'avec un degré de certitude de 0.5). Si on trouve sur le site *"ouvert tous les jours"*, il faudra alors donner une possibilité d'ouverture de 1 à tous les jours, et diminuer le degré de possibilité de fermeture à 0 pour chaque jour. L'intérêt de ce genre d'information contextuelle est avant tout de pouvoir manipuler une connaissance supposée par défaut. En pratique ce doit être l'information la moins prioritaire des informations dont on dispose.

	Lundi	Mardi	Mercredi	Jeudi	Vendredi	Samedi	Dimanche
μ_O	0.5	1	1	1	1	1	1
μ_F	1	0.5	0.5	0.5	0.5	0.5	0.5

TAB. 2 – *Connaissances contextuelles relatives à un restaurant*

5.4 Exemple complet

Dans cette partie, nous allons essayer de modéliser notre connaissance sur l'ouverture d'un restaurant. Nous connaissons les moments de la semaine où les restaurants sont habituellement ouverts (ces moments sont résumés dans la Table 2). Sur le site internet du restaurant que l'on considère ici, il est écrit : *Ouvert Juillet Août, sauf jours fériés du Mardi au Dimanche*. Nous allons traiter successivement chaque information, de la plus générale à la plus spécifique.

Dans notre exemple, nous allons commencer par discuter des jours d'ouverture possibles de la semaine. Le restaurant doit être ouvert *"du Mardi au Dimanche"*, ce qui nous permet de modifier notre information contextuelle supposée, de manière à obtenir les possibilités d'ouverture et de fermeture la Table 3.

	Lundi	Mardi	Mercredi	Jeudi	Vendredi	Samedi	Dimanche
μ_O	0	1	1	1	1	1	1
μ_F	1	0	0	0	0	0	0

TAB. 3 – *Ouvert du mardi au dimanche*

Ensuite, attachons-nous à la période *"Ouvert Juillet Août"*. Comme il n'est pas précisé explicitement ouvert *à partir du 1er* Juillet, on se doit de supposer que la date d'ouverture se situe aux alentours du 1er Juillet, ce qui donne les possibilités

d'ouverture/fermeture représentées en haut de la Figure 2, (on restreint pour plus de clarté les illustrations aux mois de Juin et Juillet). Contrairement à la Figure 1, le temps est discrétisé. Dans un souci de lisibilité nous ne considérons qu'un seul moment de la journée : le soir.

Nous pouvons maintenant agréger nos connaissances relatives aux jours de la semaine, et à la période d'ouverture. Pour cela nous avons besoin des connaissances relatives à l'année considérée. Plaçons nous dans l'année 2009, les 22 et 29 Juin 2009 ainsi que les 6, 13, 20 et 27 Juillet tombent un lundi. Enfin, nous pouvons utiliser l'information *"sauf jours fériés"* qui nous semble être la plus spécifique. Pour cela nous avons encore besoin des connaissances relatives à l'année considérée pour savoir que le 14 Juillet et le 15 août sont fériés (cela pourrait éventuellement faire partie d'un savoir valide chaque année). L'agrégation de toutes nos connaissances doit donc donner les possibilités représentées en bas de la Figure 2.

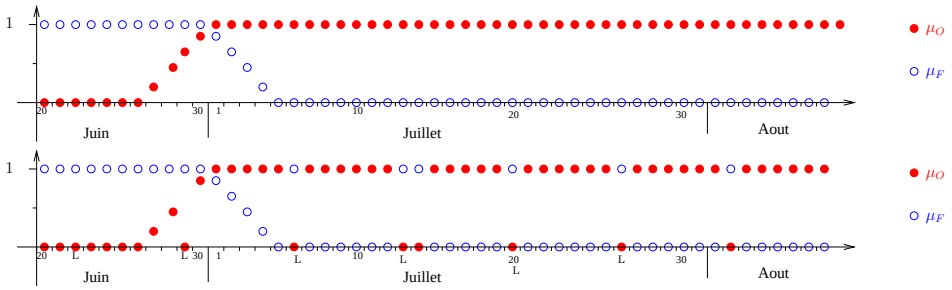

FIG. 2 – *Ouvert Juillet Août*

5.5 Exploitation des connaissances temporelles

Le schéma général d'interrogation des données temporelles utilisée dans le cadre du projet Eiffel est du type : *chercher un créneau de "n jours successifs" sur une période donnée* ou *chercher un créneau de "n jours éparpillés" sur une période donnée*. Par exemple : *"ce restaurant est-il ouvert le <une date> ?", "je cherche un hôtel pour la semaine du <samedi j> au <samedi j+7> ?", "je veux réserver une semaine au mois de Juillet", "je veux profiter de mon séjour pour visiter tel musée ?"...* Il faut également penser à la combinaison de ces requêtes temporelles. Par exemple : *"Je souhaite partir 2 semaines cet été : une semaine dans un village de vacances à la montagne où je puisse faire 2 jours de ski d'été suivie d'une semaine dans un camping de bord de mer pendant un festival d'été."*

Lors de la présentation des résultats à l'utilisateur du portail, il n'est pas envisageable de présenter deux scores, le premier étant le degré de possibilité π et le second le degré de certitude N. D'autant plus que si le degré de certitude est non nul, cela signifie que le degré de possibilité vaut 1. Inversement, si le degré de possibilité est différent de 1 cela implique que le degré de certitude vaut zéro. On peut donc agréger ces deux degrés d sur une échelle unique graduée de 0 à 100 ayant pour valeur $d = \frac{\pi+N}{2} \times 100$. Pour $d = 100$, la réponse est certaine, tandis que pour $d = 0$, la solution associée est impossible. Entre les deux, la réponse est de moins en moins probable et il est recommandé de la vérifier (par exemple en téléphonant aux établissements concernés). Cette échelle pourra donc permettre de classer les différentes réponses d'une requête. En pratique on pourra éventuellement se contenter d'afficher, s'il y en a, les réponses certaines ($d = 100$), et n'afficher les réponses incertaines que s'il n'existe pas de réponse sûre.

Supposons par exemple que la Figure 1 représente l'ouverture d'un hôtel A, et la Figure 3 l'ouverture d'un hôtel B. On pose la question *"peut-on se loger dans l'hôtel A ou B entre le 20 et le 28 février ?"*. Ici la question sous entend qu'on

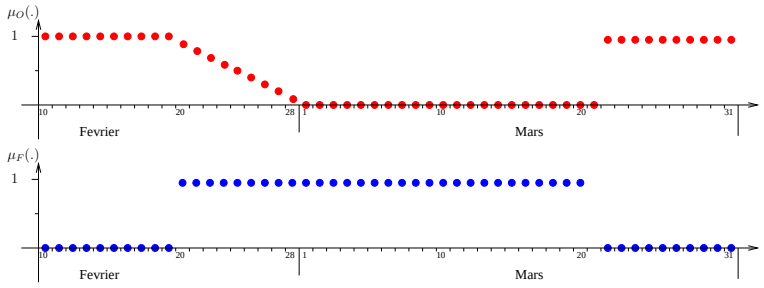

FIG. 3 – *Fermé de fin février au 21 Mars*

veut se loger tous les jours de la période en étant prêt à changer d'hôtel en cours de séjour. La réponse sera alors oui avec

un degré de possibilité $0,5$ et de certitude de 0. Ce degré de possibilité est calculé de la façon suivante : c'est le minimum du maximum des degrés de possibilité d'ouverture des établissements (i.e. le minimum de la fonction supérieure de la Figure 4 sur l'intervalle des dates recherchées). Le score présenté à l'utilisateur sera donc de 25, sur l'échelle graduée jusqu'à 100.

En revanche si l'on pose la question "peut-on se loger une nuit dans l'hôtel A ou B entre le 20 et le 28 février ?", la réponse sera oui avec un score de 100 (réponse certaine).

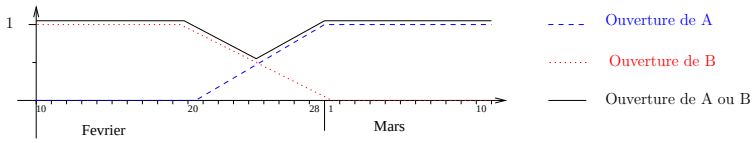

FIG. 4 – *Ouverture de A ou B*

6 Conclusion

Dans le cadre d'un projet concret d'acquisition et gestion de connaissances touristiques, nous proposons d'utiliser une modélisation possibiliste des données temporelles extraites automatiquement afin de mieux prendre en compte l'imprécision de ces données. Pour chaque objet touristique et à chaque chaque date considérée, deux degrés de possibilité sont calculés. Un degré d'ouverture et un degré de fermeture. Ainsi, le système sera en mesure de répondre aux requêtes en proposant les solutions les plus sûres, mais pourra également proposer des solutions envisageable mais non certaine sur la période demandée. Dans ce dernier cas, un degré de fiabilité de la réponse peut être calculé.

Pour l'implémentation de ce modèle possibiliste, nous étudions deux approches. La première consiste en la création et le stockage de tables représentant les distributions de possibilité d'ouverture et de fermeture. Cette opération pourra alors être effectuée dès la fin de l'analyse lexicale. Pour exécuter une requête, il suffira alors de faire un accès aux tables représentant nos connaissances. Le principal inconvénient de cette solution est l'espace mémoire consommé. Une autre solution consiste à ne garder en mémoire que le résultat des annotations brutes extraites. Il faut alors calculer à la volée pour chaque requête les degrés de possibilité d'ouverture des établissements susceptibles d'être concerné par la requête.

Références

Carloni, O., M. Leclère, et M. Mugnier (2007). Introducing reasoning into an industrial knowledge management tool. *Applied Intelligence*.

Dubois, D. et H. Prade (1988). *Possibility Theory. An Approach to Computerized Processing of Uncertainty*. Plenum Press, New York.

Noël, L., O. Carloni, N. Moreau, et S. Weiser (2008). Designing a knowledge-based tourism information system. In *Int. J. of Digital Culture and Electronic Tourisme, Special Issue on National Tourism Organisations and Exploitation of Information Technologies*. Inderscience Publishers Ltd.

Weiser, S. (2008). Informations spatio-temporelles et objets touristiques dans des pages web : repérage et annotation. In *Actes de Recital*, Avignon.

Weiser, S., P. Laublet, et J.-L. Minel (2008). Automatic identification of temporal information in tourism web pages. In E. L. R. A. (ELRA) (Ed.), *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

Summary

In the framework of a semantic e-tourism portal, data is automatically collected from web pages in order to fill a knowledge base. Amongst other types of information, temporal information is essential, concerning dates and opening times of tourism resources. These opening times are often incomplete, vague or ambiguous and their interpretation requires implicit knowledge about the tourism field. In this paper, we focus on the acquisition process and we shed light on the problems revealed by this temporal information. We suggest a possibilist modelisation of this extracted knowledge in order to take into account the vagueness of such knowledge and allow flexibility for its use.