



République Algérienne Démocratique et Populaire



Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université AMO de Bouira

Faculté des Sciences et des Sciences Appliquées

Département d'Informatique

Mémoire de Master

en Informatique

Spécialité : Ingénierie des systèmes d'informations et logiciels

Thème

Génération automatique de brochure touristique

Encadré par

— BAL KAMEL

Réalisé par

— MAHFOUD SALMA

2017/2018

Remerciements

Avant tout je remercie mon Dieu le tout puissant d'avoir me donner la force et le courage de mener à terme le présent travail.

Je tiens à exprimer mes remerciements avec un grand plaisir et un grand respect à mon encadreur, Monsieur BAL Kamel pour ses conseils, pour sa patience, sa disponibilité, son sérieux, ses remarques, ses conseils, son respect et sa bienveillance, qui m'ont permis de réaliser ce travail dans les meilleures conditions.

Mon gratitude va également aux membres de jury pour avoir juger et évaluer mon travail.

A nos chers enseignants du département d'informatique, un remerciement particulier et sincère pour tous les efforts que vous avez fournis pour nous encadrer tout au long de ces années.

Je ne peux pas nommer ici toutes les personnes qui de près ou de loin m'ont aidé et encouragé mais je les remercie vivement.

Dédicaces

Je dédie ce travail à mon cher père SALEM et ma chère mère DAOU, que nulle dédicace ne puisse exprimer mes sincères sentiments, pour leur patience illimitée, leurs encouragements contenus, leurs prière pour moi, leurs aides, leurs tendances inestimables.

Je suis très reconnaissante pour tous les sacrifices que vous n'avez cessé de me donner depuis ma naissance, durant mon enfance et mes études du primaire jusqu'à l'université.

Puisse Dieu, le tout puissant, vous garde et vous procure santé, longue vie et bonheur. A mes chers frères ADEL et sa femme ZOHRA et ses petites anges MARAM et KOUSSAY. A DJAMEL et sa femme AHLAM et son ange FAROUK, à mes frères KHALED et ADAM pour leurs conseils et soutiens. A mes deux chères sœurs AKILA et son mari FAYCEL et sa belle petite RANDA, à SAMIRA et son mari TOUFIK et ses anges enfants ABD EL-RAHMANE et GHOFRANE pour leurs conseils et soutiens. A mes chères copines de ma vie SABRINA, ASMA, FELLA, FATIMA, à toutes mes amis de groupe ISIL.

MAHFOUD Selma

Table des matières

Table des matières	i
Table des figures	ii
Liste des abréviations	iv
Introduction générale	1
0.1 Contexte et problématique :	1
0.2 Objectif	2
0.3 Organisation du mémoire :	2
1 La recherche d'informations classique	4
1.1 Introduction :	4
1.2 Définition et les notions de base de la recherche d'information :	5
1.3 Processus de la RI :	7
1.3.1 Indexation :	8
1.3.2 Appariement document-requête :	11
1.3.3 Reformulation de requête :	12
1.4 les modèles de recherche d'information :	12
1.4.1 Le modèle booléen :	13
1.4.2 Le modèle vectoriel :	13
1.4.3 Le modèle probabiliste :	15
1.5 Evaluation des Système de recherche d'information SRI :	17
1.5.1 Rappel et Précision :	17
1.5.2 La collection TREC :	18

1.5.3	Autres campagnes :	18
1.6	Conclusion :	19
2	La recherche d'informations agrégée	20
2.1	Introduction :	20
2.2	Qu'est ce que la recherche d'informationa grégée?	21
2.2.1	Définition :	21
2.2.2	Le processus générique de la recherche d'information agrégée : . . .	21
2.2.3	Structure d'agrégat :	23
2.3	les problématiques liées à la RI agrégée :	24
2.4	Les techniques d'agrégations et les approches d'agrégations :	25
2.4.1	L'approche d'agrégation par clustering :	25
2.4.2	L'approche d'agrégation par résumé multi-documents :	26
2.4.3	La génération d'un document à partir de plusieurs documents : . .	27
2.4.4	L'approche d'agrégation relationnelle :	27
2.4.5	Les vues agrégées « Aggregated view » :	27
2.5	Conclusion :	31
3	Brochure touristique	32
3.1	Introduction :	32
3.2	L'e-tourisme :	33
3.2.1	Du Web 1.0 au Web 3.0 :	33
3.2.2	Le Web au service du tourisme en ligne :	33
3.2.3	Qu'est ce que le e-tourisme?	34
3.3	Les brochures touristiques :	34
3.3.1	Définition :	34
3.3.2	Les objectifs et les intérêts d'une brochure touristique :	34
3.3.3	Contenu et structure d'une brochure touristique :	36
3.3.4	Synthèse de l'étude de la brochure :	38
3.3.5	Limitations des brochures touristiques classiques :	39
3.4	Conclusion :	40
4	Conception et modélisation	42
4.1	Introduction	42

4.2	La conception de notre brochure :	42
4.3	Présentation générale du système :	43
4.3.1	Le processus de génération de la brochure :	44
4.3.2	Les sources d'informations utilisées :	44
4.3.3	Dispatching	45
4.3.4	Sélection des nuggests :	45
4.3.5	L'agrégation des résultats :	48
4.4	La modélisation UML (Unified Modeling Language) de système :	49
4.4.1	Les objectifs d'UML :	49
4.4.2	L'architecteur de notre système :	50
4.4.3	Modélisation du système	50
4.5	Conclusion :	54
5	Réalisation et implémentation	55
5.1	Introduction	55
5.2	Environnement de travail	55
5.2.1	Environnement matériel	55
5.2.2	Environnement logiciel et développement	55
5.3	Présentation de l'application	59
5.3.1	La page d'accueil :	59
5.3.2	La brochure	59
5.4	Conclusion :	69
	Conclusion générale	70
	Bibliographie	72

Table des figures

1.1	processus U de recherche d'information	7
1.2	les fomules de TF-IDF	11
1.3	une représentation du modèle vectoriel avec deux documents et une requête	14
1.4	les mesures appliquées dans le modèle probabilistes	16
2.1	le schéma conceptuel pour le processus de recherche agrégée	22
2.2	exemple de yippy	26
2.3	exemples de recherche relationnelle agrégée sur Google Squared	28
2.4	exemple Google :design blended	29
2.5	exemple de alphaYahoo(design non-blended)	30
3.1	les sources des information utilisées après avoir quitté le domicile	35
3.2	brochure Amsterdam	37
3.3	la section de restaurants à Amsterdam	39
4.1	le contenu visuel et textuel des notre brochure	43
4.2	Le processus de génération de notre brochure	44
4.3	identification des sources	45
4.4	le dispatching de la requête «Taghit»	46
4.5	la sélection des nuggets	47
4.6	la structure de la brochure	48
4.7	la structure de la galerie d'images	49
4.8	Les diagrammes de modlisation par UML	50
4.9	l'architecture de notre système	51

4.10	Diagramme des cas d'utilisation général du système	52
4.11	Diagramme de séquence pour Imprimer la brochure	53
4.12	Diagramme de classes	53
5.1	Page d'accueil	59
5.2	Page récapitulant le contenu	60
5.3	Description du lieu	60
5.4	Page des Hôtels	61
5.5	Page des restaurants	62
5.6	Page des événements	63
5.7	Page de la carte géographique	64
5.8	Page des mosquées	64
5.9	Page des musées	65
5.10	Page des météo	65
5.11	Page des stations de transport	66
5.12	Page des services supplémentaires	67
5.13	Page de galerie d'images	68

Liste des tableaux

1.1	les mesures appliquées dans le modèle vectoriel	15
-----	---	----

Liste des abréviations

API Application Programming Interface

IHM Interface Homme Machine

Introduction générale

0.1 Contexte et problématique :

Au début des années 2000, le monde a connu l'explosion de la bulle Internet, ce qui a conduit à la diffusion de la technologie et des informations qui permet à l'internet étant devenu un chiffre difficile dans tous les domaines et progressivement investi ce réseau pour but de développer et généraliser l'informatique; cette révolution a généré une énorme quantité d'informations sur le Web sur des milliards de pages. Il est donc très difficile de trouver ce que l'utilisateur recherche sous cette masse d'informations. La recherche d'information (RI) est une branche informatique qui s'intéresse à la mise en place d'outils automatisés permettant à la facilité d'accès à des informations précises et pertinent grâce à un processus de recherche d'information qui est fournit à mettre en relation l'ensemble des informations disponibles d'une part et les besoins de l'utilisateur d'une autre part , et les besoins en information des utilisateurs d'autre part.

À l'heure actuelle, le développement technologique considérable sur le Web a affecté l'évolution des besoins des utilisateurs, car il ne s'agit pas d'informations suffisantes extraites de différents moteurs de recherche, qui reposent sur une source unique d'informations, mais il souhaite des informations différentes types d'une multitude de sources d'information afin de mieux répondre et satisfait les besoins d'utilisateur. C'est dans ces notions que la recherche d'information agrégée est apparue. Notre objectif de travail consiste à développer un système de génération automatique de brochures touristiques en appliquant le paradigme de recherche d'information agrégée. la génération auto de brochures touristiques est un bon exemple d'application de la RI agrégée car on doit faire appel aux deux notions principal de la RI agrégée (multi source et agrégation) pour

générer une brochure touristique :

- Une brochure est composé d'information touristique est une information variées des différents types et hétérogénies (texte, images, données structurées, ..).
- On a besoin donc de sélectionner des données issues des plusieurs sources pour générer une brochure.
- fusion ou agrégation des résultats des données sélectionnées pour enrichir le contenu de la brochure touristique .

La masse d'information volumineuse dans le web pose plusieurs issues pour le processus de recherche d'information agrégée, il y'a quelques problématiques de base :

- la problématique de Sélection des sources (quelles sources utiliser pour d'offrir les informations Pertinentes une requête donnée) .
- Le problème du choix des résultats à sélectionner depuis chaque source sachant que chaque source peut restituer des milliers de résultats.

0.2 Objectif

: Le travail présenté dans ce mémoire a pour objectif principal de concevoir et développer un système de génération automatique de brochures touristiques en s'inspirant du paradigme de recherche d'information agrégée. Ce système doit servir comme un outil fiable aux touristes pour la découverte des zones touristiques à travers le monde.

0.3 Organisation du mémoire :

Notre mémoire doit être structuré comme suit : Dans le chapitre 1, La recherche d'information : nous donnons les concepts de base de la recherche d'information. Qui sont les notions de besoin en information, de requête, de document et de pertinence et le processus d'indexation. En suite nous décrivons les différents modèles de la recherche d'information en particulier le modèle booléen, le modèle vectoriel et le modèle probabiliste. Et dernier point concernera l'évaluation des systèmes de recherche d'information. Dans le chapitre 2, La recherche d'information agrégée : nous donnons un aperçu sur les principaux travaux de recherche menés pour l'application du principe d'agrégation dans le contexte de la RI. D'abord nous allons présenter les concepts de base, ainsi que le processus générique de la recherche d'information agrégée. Après, montrer les problématiques liées à la RI

agrégée. Enfin, citer les différentes approches de la recherche d'information agrégée. Dans le chapitre 3, Brochure touristique, nous traitons la notion du e-tourisme, la brochure touristique, ses objectifs et ses intérêts, Ensuite une étude d'une brochure touristique afin d'identifier son contenu. Pour le chapitre 4, La conception et modélisation, nous présentons la conception de notre solution et l'organisation de la brochure, ainsi que la modélisation du système. Terminer avec le chapitre 5, L'implémentation, décrire le coté technique du travail, on présente les différents outils et techniques qu'on a utilisé pour concrétiser le projet, et on montre à la fin des captures d'écrans du système.

La recherche d'informations classique

1.1 Introduction :

Dans ces dernières années le monde produit un volume important d'information, ces informations sont disponibles partout, l'accroissement de la masse gigantesque des informations produite, provoque une difficulté de la gestion et de l'organisation de ces informations. C'est pour cela des études étaient entretenues pour établir des processus permettant d'accéder à ces informations, non seulement n'importe information mais celle qui répond aux besoins de la recherche des utilisateurs. Pour cela il existe plusieurs méthodes et techniques pour l'acquisition, l'organisation, le stockage, la recherche et la sélection de l'information pertinente pour un utilisateur, Lors de web, la recherche d'information (RI) peut être définie comme un ensemble des méthodes et techniques dont la finalité est de délivrer un ensemble de documents à un utilisateur en fonction de son besoin en informations pertinentes. Le défi est de pouvoir, parmi le volume important de documents disponibles trouvé ceux qui correspondent au mieux à la demande de l'utilisateur. L'application de la RI (recherche d'information) nécessite de développer des systèmes automatisés efficaces permettant de collecter, organiser, rechercher et sélectionner des informations pertinentes répondant à des besoins utilisateurs, qui sont les systèmes de recherche d'information (SRI) L'objectif de ce chapitre est de donner un aperçu sur les principes de la recherche d'informations classique qui est basée sur les concepts de base de RI, ainsi que les différents modèles de la recherche d'informations en particulier le modèle booléen, le modèle vectoriel et le modèle probabiliste, le dernier point à traiter concernera l'évaluation des systèmes de recherche d'information.

1.2 Définition et les notions de base de la recherche d'information :

On distingue plusieurs définitions de la RI dans tout ces domaines d'intérêt, nous citons quelques définitions :

« La recherche d'information est une activité dont la finalité est de localiser et de délivrer des granules documentaires à un utilisateur en fonction de son besoin en informations. » [1]

« la recherche d'information est l'opération qui permet à partir d'une expression des besoins en information d'un utilisateur de retrouver l'ensemble des documents contenant l'information recherchée » [2]

« La recherche d'information est une branche de l'informatique qui s'intéresse à l'acquisition, l'organisation, le stockage, la recherche et la sélection d'information » [3]

« La recherche d'information est une discipline de recherche qui intègre des modèles et des techniques dont le but est de faciliter l'accès à l'information pertinente pour un utilisateur ayant un besoin en information » [4]

Toutes ces définitions partagent la même idée qui concerne l'objectif de la RI qui est la sélection dans une collection d'informations (items, documents,..), les informations pertinentes répondant à des besoins d'utilisateur. [5] D'après ces définitions on peut Extraire les notions centrales et de base de la RI :

Besoin d'information : Le besoin d'information est une expression mentale d'un utilisateur.

Requête : la requête constitue l'expression du besoin en information de l'utilisateur. Elle représente l'interface entre le SRI et l'utilisateur. Divers types de langages d'interrogation sont proposés dans la littérature. Une requête est un ensemble de mots clés, mais elle peut être exprimée en langage naturel, booléen ou graphique. [6]

Document : le document est un ensemble des informations représentées par des termes ou mots ,sous une forme permanente lisible, exploitable et accessible par le SRI.

Pertinence : selon Goffman, 1969 : «... la pertinence de l'information d'un document dépend de ce qui est déjà connu sur le sujet, et à son tour affecte la pertinence d'autres documents examinés ultérieurement » Pour être en mesure d'offrir aux utilisateurs les informations répondant le mieux à leurs besoins, tout système de recherche d'information s'appuie sur un modèle de calcul de pertinence qui, pour chaque requête, calcul le score de pertinence de chaque donnée (document). Celles qui auront le meilleur score de pertinence seront présentées à l'utilisateur.. [7] Donc la notion de pertinence au cœur de tout système de recherche d'information, elle permet de lier le jugement selon l'utilisateur (pertinence utilisateur) et évaluation des systèmes de RI (pertinence système) . D'où l'objectif de tout SRI est de rapprocher la pertinence système de la pertinence utilisateur .Alors on distingue deux types de pertinence :

la pertinence utilisateur : correspond à la satisfaction de l'utilisateur par apport à l'ensemble des documents restitués par le SRI. (pertinence subjective, cognitive)

- **pertinence subjective :** venant de l'usager dans la mesure où il permet à ce dernier de sélectionner une information qui non seulement lui apporte un élément de connaissance nouveau mais qui soit aussi compréhensible et utilisable par lui.
- **pertinence cognitive :** relation entre l'état de la connaissance de l'utilisateur et l'information sélectionnée [8]

la pertinence système : le système mesure un degré de pertinence, une valeur de similitude entre un document et une requête. (pertinence algorithmique, objective)

- **pertinence algorithmique :** basés sur la concordance entre une requête et les termes indexés d'un document, est la plus facile à décrire [9]
- **pertinence objective :** Le caractère objectif de la pertinence se distingue par une conception ancrée dans une tradition nettement positiviste, que ce soit par une définition algorithmique(ou calculatoire) ou par adéquation objective entre une requête et le contenu d'un document. [9]

1.3 Processus de la RI :

Pour effectuer d'une façon efficace la satisfaction de l'utilisateur la RI constitue un processus contenant plusieurs étapes et tâches, ce processus se compose de trois fonctions principales :

- L'indexation.
- L'appariement document-requête
- Reformulation de requête

Le processus est schématisé par un processus en U Le déroulement de ce processus com-

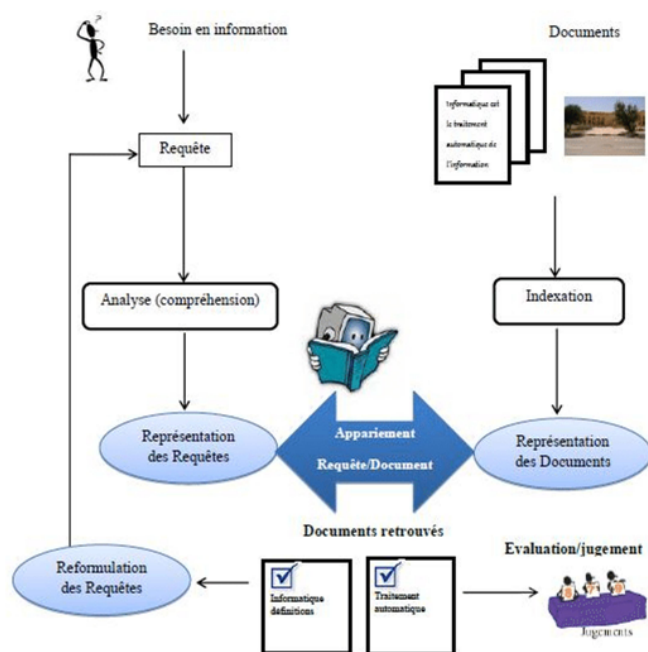


FIGURE 1.1 – processus U de recherche d'information

mence par un besoin d'information de l'utilisateur qui se traduit sous forme de requête , cette requête est soumise à un système de recherche d'information (SRI) ,le SRI doit réaliser l'indexation des documents pour les représenter , il analyse la requête dans le même but, en se basant sur un modèle de RI, il fait l'appariement requête/document pour collecter tout document répondant à la requête, sinon il propose une reformulation de la requête. Les Résultats sont soumis à l'évaluation et le jugement des utilisateurs

1.3.1 Indexation :

La définition proposée par l'AFNOR en 1993, est la suivante : « l'indexation est le processus destiné à représenter par les éléments d'un langage documentaire ou naturel des données, résultat de l'analyse du contenu d'un document ou d'une question ».

Processus d'indexation : c'est la transformation du document et de la requête en une représentation informatique qui reflète son contenu informationnel. Le résultat de l'indexation est un descripteur pour chaque document. Le plus souvent pour un document, ce descripteur contient une liste de termes auxquels sont associés des poids, qui tentent de caractériser le degré de représentativité de ces termes dans le document. Le but général de l'indexation est d'identifier l'information contenue dans tout texte et de le représenter au moyen d'ensemble appelé index pour permettre la comparaison entre la représentation d'un document et d'une requête. L'indexation peut être manuelle, semi-automatique et automatique et aussi basé sur deux vocabulaires contrôlé et libre :

- **Manuelle :** assure une haute pertinence à l'aide d'un vocabulaire contrôlé Prédéfini, l'indexation manuelle est une opération réalisée par un documentaliste, l'inconvénient de cette méthode est la possibilité d'avoir deux listes différentes de descripteurs pour un même document si le travail est fait par deux indexeurs différents.
- **Semi-automatique :** Un premier processus automatique permet d'extraire les termes du document. Cependant le choix final reste au spécialiste du domaine ou au documentaliste pour établir les relations entre les mots clés et choisir les termes significatifs, et ce grâce a une interface interactive. [10]
- **Automatique :** est celle qui a été la plus étudiée en recherche d'information. Il s'agit d'automatiser complètement la procédure d'indexation. On y distingue : l'extraction automatique des termes, l'utilisation d'un anti-dictionnaire pour éliminer les mots vides, la lemmatisation, le repérage des groupes de mots, la pondération des mots avant de créer l'index, etc. Le résultat de l'indexation est un ensemble de termes définissant ce qu'on appelle le langage d'indexation. [11]
- **Vocabulaire contrôlé :** un ensemble de concepts bien défini est assigné à chaque document manuellement ou d'une manière automatique , Il est représenté sous plusieurs formes :lexique/thesaurus/ontologie/réseau sémantique.
- **Vocabulaire libre :** les concepts sont extraits automatiquement des documents.

Les étapes principales du processus d'indexation :

Tokénisation / segmentation : Consiste à transformer un texte d'un document en ensemble des termes qui est une unité lexicale, cette étape permet de reconnaître les séparations, les chiffres, les mots, les ponctuations; cette unité est utilisée lors de la recherche.

Élimination des mots vides : on utilise souvent un anti-dictionnaire appelé aussi stoplist, liste qui contient les mots trop fréquents mais pas utiles et ne doivent pas être inclus dans l'index. Cette étape consiste à supprimer les mots vides (pronoms, des conjonctions. . . .) et permet une réduction de l'index; à la fin on garde un ensemble des mots qui sont considérés comme des index;

Lemmatisation / radicalisation (racinisation) : Processus morphologique permettant de regrouper les variantes d'un mot; chaque mot appartient à une catégorie morphologique (flexionnel / dérivationnel)

- le lemme; la forme canonique d'un mot; s'obtient par la morphologie flexionnelle
La lemmatisation regroupe les différentes formes que peut revêtir un mot, soit : le nom, le pluriel, le verbe à l'infinitif, ..etc. par exemple : j'ai , tu as Le lemme est un verbe « avoir »;
- la racine : c'est la nature d'un mot s'obtient par la morphologie dérivationnel .
par exemple : économie, économiquement, économiste La racine « économ ».
la lemmatisation Une technique qui repose sur plusieurs règles de transformations pour conduire permet ces règles on a : **règle de type** : condition -i action On a par exemple si mot se termine par 's' alors supprimer la terminaison grâce à des algorithmes. L'algorithme le plus connu est l'algorithme de « Porter ».

analyse grammatical : qui permet d'identifier la nature et la fonction des mots. Deux techniques utilisées : Utilisation de lexique (dictionnaire). Tree-tagger (gratuit sur le net).

Troncature : Tronquer les mots à X caractères et plutôt les suffixes Exemple troncature à 7 caractères : économiquement : écomoni

Fichier inverse : un mécanisme orientée mots (termes) pour l'indexation du document (collection de texte) afin de faciliter et accélérer la tâche de recherche .Le fichier

inverse est structurée en deux parties : le dictionnaire (ou vocabulaire) : contenant tout les termes du vocabulaire ; les occurrences (posting) : contenant pour chaque terme du dictionnaire ,la liste des documents le contenant . le posting peut être enrichi avec d'autres d'informations comme : la fréquence des termes ,position des termes,

Pondération : La pondération des termes caractérise l'importance des termes dans un document ,on effectue à chaque terme de l'index un poids qui mesure son importance dans les documents qui les contiennent, pour ce traitement il existe plusieurs méthodes statistiques basée sur le nombre d'occurrence (la fréquence); la fréquence des termes correspond à l'occurrence de chaque terme dans un ou plusieurs documents. On a plusieurs méthodes statistiques basé sur la loi de Zipf et TF*IDF pour la pondération.

- **la loi de Zipf (George Kingsley Zipf) :** Le principe de la loi est "principe du moindre effort" :il est plus simple pour un auteur (rédacteur d'un document) de répéter les mots que d'en chercher de nouveaux. [12].La loi zipf est une observation empirique concernant la fréquence des mots dans un texte [12]. $ft = k \div rt$
 ft c'est fréquence (nombre d'occurrence du terme t),
 rt c'est le rang (à base de fréquence) du terme t ,
 k c'est une constante (spécifique à la collection)
- pondération TF x IDF : c'est la plus utilisée des méthodes de pondération basée sur la combinaison de deux facteurs tf et idf et dépend aussi du modèle de RI.

TF (Term Frequency) : la fréquence du terme dans un document (pondération locale) ,le TF est utilisé selon plusieurs déclinaisons : $tf = \text{freq}(t,d)$, $tf = 1 + \log(\text{freq}(t,d))$.
 freq c'est la fréquence du terme t dans un document d .

IDF (Inverse of Document Frequency) : la fréquence du terme dans la collection des documents (pondération globale) , il est représenté par la formule suivante : $idf = \log(N \div nt)$.

N = le nombre de documents dans la collection

nt = le nombre de documents contenant t (terme)

le TF-IDF (TermFrequency-Inverse Document Frequency) : cette méthode permet de combiner les deux pondérations locale et globale qui déterminent l'évaluation de l'importance d'un terme contenu dans un document, relativement à une collection. Le poids $w(t,d) = tf * idf$ consiste à quantifier l'importance de terme t pour décrire le contenu du document d , les formules possibles pour le calcul de poids $w(t,d)$ sont les suivantes :

$$w(t,d) = tf * idf = \left\{ \begin{array}{l} \frac{(1 + \log(freq(t,d))) * \log \frac{N}{n_t}}{\sum_{t' \in d} (1 + \log(freq(t',d))) * \log \frac{N}{n_{t'}}} \\ \frac{freq(t,d)}{k1.(1-b + b * \frac{dl}{avgdl}) + freq(t,d)} * \log \frac{N - n_t}{n_t} \end{array} \right.$$

FIGURE 1.2 – les formules de TF-IDF

N = le nombre de documents dans la collection

n_t = le nombre de documents contenant t (terme)

dl = longueur du document (nombre de termes)

$K1$, b = des constantes

1.3.2 Appariement document-requête :

Ce processus permet d'effectuer la comparaison entre la représentation de document et celle de la requête, liée aux opérations d'indexation et de pondération des termes de la requête et des documents. La comparaison revient à calculer un score représentant la pertinence du document vis-à-vis de la requête, ce poids calculé à partir d'une fonction ou d'une probabilité de similarité $RSV(Q,d)$ (Retrieval Status Value), où Q est une requête et d un document, cette fonction d'appariement appartenant à un système de RI permet d'ordonner et classer les documents renvoyés à l'utilisateur ainsi qu'elle traduit le degré de pertinence des résultats, on retrouve deux types d'appariement exact et approché.

Appariement exact : Les Requêtes spécifient de manière précise les critères recherchés et l'ensemble des documents respectant exactement la requête sont sélectionnés, mais pas ordonné.

Appariement approché : Les Requêtes décrivent les critères recherchés dans un document Et les documents sont sélectionnés selon un degré de pertinence (similarité/ probabilité) vis-à-vis de la requête et sont ordonnés .

1.3.3 Reformulation de requête :

Des fois c'est difficile a l'utilisateur de formuler son besoin d'information exact en requête ,ce pendant les résultats retournés par le SRI ne lui conviennent parfois pas.

La reformulation de la requête est un processus ayant pour objectif de générer une requête plus ciblée et adéquate à la recherche d'information, que celle initialement formulée par l'utilisateur. Ce processus consiste généralement à modifier la requête initiale de l'utilisateur via plusieurs techniques de classifications et modification .afin de retourné les documents plus pertinents à requête de l'utilisateur.

1.4 les modèles de recherche d'information :

Les modèles de recherche d'information ou modèles d'appariement document /requête au objectif de fournir un processus de RI .ils présentent un cadre théorique pour la modélisation de la mesure de pertinence. De très nombreux modèles sont proposés dans la littérature ,nous allons décrire les principaux modèles proposés pour la RI structurée ,qui sont étroitement liés, ces modèles de RI se distinguent par le principe d'appariement .on peut distinguer trois catégories de modèles de RI :

1. les modèles basés sur la théorie des ensembles : c'est le modèle booléen, il représente la requête sous forme une expression logique et les operateurs logique qui séparent ces termes .
2. les modèles algébriques : le modèle vectoriel, qui fournit la pertinence d'un document/ requête définie par des mesures de distance dans un espace vectoriel.
3. les modèles probabilistes :qui se base sur la théorie des probabilités, le modèle probabliste fournit une pertinence d'un document/ requête qui est définie comme une probabilité

de pertinence.

1.4.1 Le modèle booléen :

Le premier est le plus simple modèle de RI, le modèle booléen est basé sur la théorie des ensembles et l'algèbre de Boole, les documents sont représentés en un ensemble de termes liés entre eux par la conjonction logique exemple : $d = t_1 \wedge t_2 \wedge \dots \wedge t_n$

Et la requête est une expression logique, les termes d'indexation sont reliés par des opérateurs booléens (logique) : AND (\wedge) OR (\vee), NOT (\neg) exemple :

$$q = t_1 \wedge t_2 \vee t_3$$

Le modèle booléen se distingue par le principe d'appariement exact basé sur la présence ou l'absence des termes de la requête dans les documents appariement $(d, q) = R(d, q) = 1$ ou 0 . En conséquence, les poids des termes dans l'index sont binaires, c'est à dire $w_{i,j}$ soit 0 , soit 1 . La fonction de correspondance de similarité $R(d, q)$ entre une requête et un document est déterminée de la façon binaire soit 1 ou 0 .

Ce modèle présente plusieurs avantages et inconvénients ; Parmi les avantages : simple et facile à mettre en œuvre, la clarté conceptuelle des systèmes booléens. Les inconvénients : La sélection d'un document est basée sur une décision binaire donc pas d'ordre pour les documents sélectionnés, Problème de collections volumineuses : le nombre de documents retournés peut être considérable. Le besoin d'information doit être traduit en expression booléenne, que la plupart des utilisateurs trouvent difficile. A cause de ces inconvénients les chercheurs proposent deux autres modèles

1.4.2 Le modèle vectoriel :

modèle vectoriel, proposé au début des années 70 par Gérard SALTON [14]. Ce modèle est un modèle algébrique où la représentation des documents et les requêtes par des vecteurs dans l'espace vectoriel engendré par tous les termes de la collection de documents et chaque terme est une dimension ; Le traitement d'une requête est alors basé sur la comparaison des vecteurs documents et requête. ce modèle se distingue par le principe d'appariement approché qui consiste à sélectionner et ordonner les documents selon un degré de pertinence (similarité) vis-à-vis de la requête. dans le modèle vectoriel on représente un document d par un vecteur de n dimension $Doc_i = (w_{i1}, w_{i2}, \dots, w_{im})$ pour $i = 1,$

$2, \dots, m$. Où w_{ij} est le poids non binaire exprimé par la calcul de $(TF*IDF)$ du terme t_j dans le document Doc_i pour indexer les termes dans les requête et dans les documents. m est le nombre de documents dans la collection,

n est le nombre de termes dans les documents de la collection.

Cette représentation reste la même pour la requête $q_k = (w_{k1}, w_{k2}, \dots, w_{km})$. Où w_{kj} est le poids de terme t_j dans la requête q_k .

une représentation vectoriel de deux vecteurs documents D_1 et D_2 et un vecteur Requête Q dans un espace associé à trois termes T_1, T_2, T_3

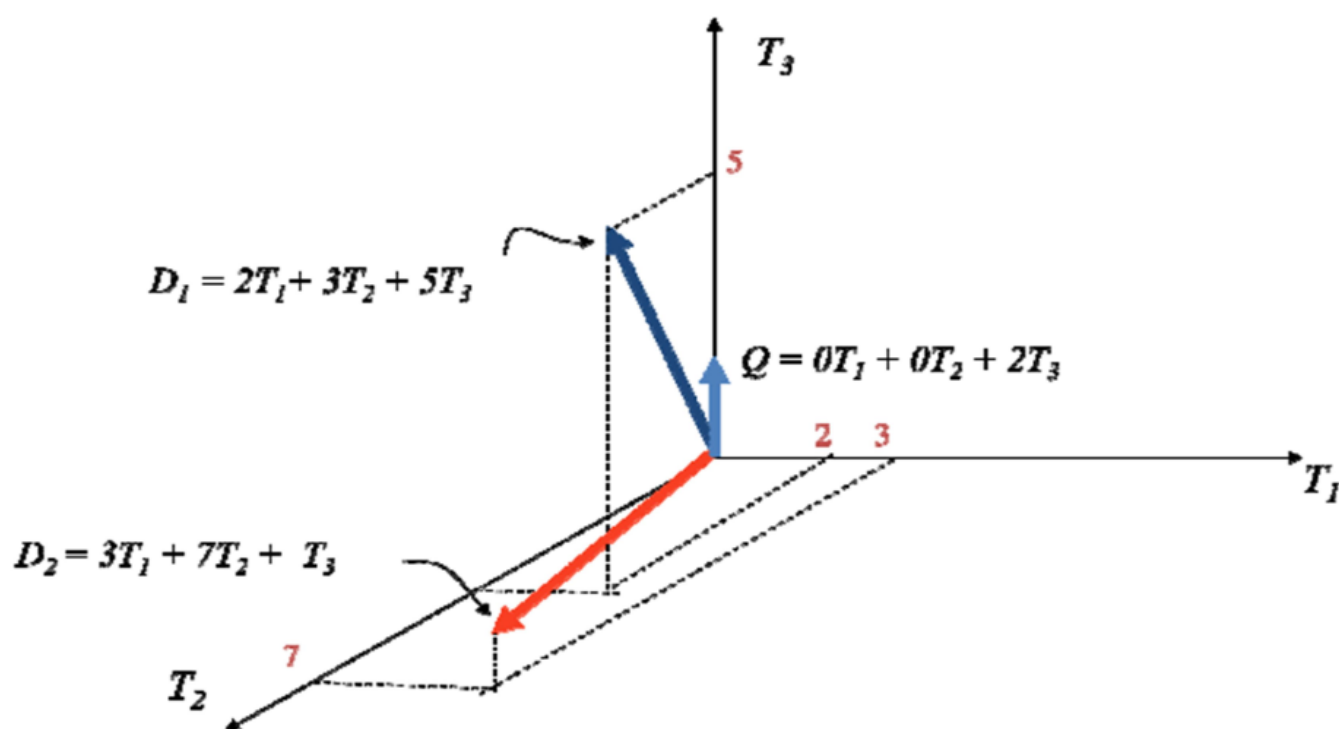


FIGURE 1.3 – une représentation du modèle vectoriel avec deux documents et une requête

Dans l'exemple de la figure, on a le vecteur de document D_1 plus similaire au vecteur de la requête Q que le D_2 ; Alors la fonction de correspondance de modèle vectoriel basée sur la similarité qui représente la degré de pertinence est traduite en une similarité vectorielle où le vecteur de document le plus similaire au vecteur de la requête est le plus pertinent. Pour déterminer la degré de similarité entre le document et la requête, on calcule la fonction de correspondance, cette fonction peut être quantifiée par plusieurs approche. nous citons quelque approche les plus utilisées; On suppose que :

Mesures	Formules
Le produit scalaire	$R(D_i, Q_k) = \sum_{j=1}^m W_{ij} * W_{kj}$
La mesure de Cosinus	$R(D_i, Q_k) = \text{Cos}(D_i, Q_k) = \frac{\sum_{j=1}^m W_{ij} * W_{kj}}{\sqrt{\sum_{j=1}^m W_{kj}^2 * \sum_{j=1}^m W_{ij}^2}}$
La mesure de Dice	$R(D_i, Q_k) = \frac{2 * \sum_{j=1}^m W_{ij} * W_{kj}}{\sum_{j=1}^m W_{ij}^2 + \sum_{j=1}^m W_{kj}^2}$
La mesure de Jaccard	$R(D_i, Q_k) = \frac{\sum_{j=1}^m W_{ij} * W_{kj}}{\sum_{j=1}^m W_{ij}^2 + \sum_{j=1}^m W_{kj}^2 - \sum_{j=1}^m W_{ij} * W_{kj}}$

TABLE 1.1 – les mesures appliquées dans le modèle vectoriel

M le nombre des termes de la collection

$D_i = (w_{i1}, w_{i2}, \dots, w_{im})$ le document

$Q_k = (w_{k1}, w_{k2}, \dots, w_{km})$ la requête

Les résultats de l'appariement est une liste des documents ordonnés par ordre de degré de pertinence décroissante selon la similarité entre le document et la requête. L'inconvénient majeur de modèle vectoriel la représentation vectorielle c'est qu'il suppose l'indépendance entre termes.

1.4.3 Le modèle probabiliste :

Le modèle probabiliste a pour objectif d'ordonner les documents selon leurs degré de probabilité de pertinence vis-à-vis la requête. Le modèle probabiliste est fondé sur le calcul de la probabilité de pertinence d'un document pour une requête. Le principe de base consiste à retrouver des documents qui ont en même temps une forte probabilité d'être pertinents, et une faible probabilité d'être non pertinents, et pour cela on cherche à estimer la probabilité qu'un document soit pertinent par rapport à une requête. [13] PERT et NPERT représentent la pertinence et la non-pertinence (ou documents pertinents et l'ensemble de documents non pertinents).

Ce modèle estime le classement des documents en diminution de probabilité de pertinence : $P(\text{PERT}—D)$ ou $(P(\text{PERT}=1—D))$ De la même façon la probabilité non pertinence $P(\text{NPERT}—D)$ ou $P(\text{NPERT}=0—D)$ On observe la pertinence ou la non pertinence sa-

chant le document D selon la présence (PERT=1) ou l'absence (NPERT=0) si le terme dans le document et la requête. Un document est sélectionné si $P(\text{PERT}|\text{D})$ plus grand que $P(\text{NPERT}|\text{D})$.

Les documents doivent être triés et classés selon une fonction de correspondance. La fonction d'appariement entre le document D et la requête Q est déterminée de la façon suivante :

$$\text{RSV}(\text{D}, \text{Q}) = \frac{P(\text{PERT}|\text{D})}{P(\text{NPERT}|\text{D})}$$

Avec :

$$P(\text{PERT}|\text{D}) = \frac{P(\text{D}|\text{PERT}) \cdot P(\text{PERT})}{P(\text{D})} \text{ et}$$

$$P(\text{NPERT}|\text{D}) = \frac{P(\text{D}|\text{NPERT}) \cdot P(\text{NPERT})}{P(\text{D})}$$

P(PERT) : probabilité de la pertinence d'un document

PERT(D|PERT) : probabilité que D fasse partie de l'ensemble des documents pertinents

P(D) : probabilité que D soit choisi

PERT(D|NPERT) : probabilité que D fasse partie de l'ensemble des documents non pertinents

P(NPERT) : probabilité de la non pertinence d'un document

FIGURE 1.4 – les mesures appliquées dans le modèle probabilistes

1.5 Evaluation des Système de recherche d'information SRI :

La RI doit garantir la satisfaction des besoins d'information de l'utilisateur ,C'est-à-dire les systèmes de recherche d'information permet aux utilisateurs deux objectifs principaux :de retrouver tous les documents pertinents pour une requête utilisateur, et de rejeter tous les documents non pertinents. L'évaluation des performances d'un SRI consiste à identifier les critères d'évaluation permettant de quantifier la performance d'un SRI. (Cleverdon, 1970) a proposé six principaux critères d'évaluation de la performance d'un SRI : (1) la couverture du discours de la collection, (2) le temps de réponse, (3) la présentation des résultats, (4) l'effort de l'utilisateur pour récupérer de l'information pertinente, (5) la précision (6) le rappel du SRI. Parmi ces facteurs, les deux derniers sont liés aux modèles de représentation de l'information du SRI. [15]

1.5.1 Rappel et Précision :

Ce sont les critères de comparaison entre les SRI les plus important qui doivent être mesurés. La réponse d'un système pour une requête avec les réponses idéales permet d'évaluer deux mesure statistiques le taux de rappel(La capacité d'un système à sélectionner tous les documents pertinents de la collection) et le taux de précision(La capacité d'un système à sélectionner que des documents pertinents).

Rappel :

Le rappel mesure la capacité du système à sélectionner tous les documents pertinents pour une requête Q, il est calculé comme suit : s pour une requête Q, il est calculé comme suit :

$$rappel = (p \wedge s) \div p$$

P : documents pertinents.

S : documents sélectionnées.

Précision :

détermine la capacité d'un SRI à sélectionner que des documents pertinents pour une requête Q , il est calculé comme suit : $rappel = (p \wedge s) \div s$

P : documents pertinents.

S : documents sélectionnées.

Lien entre Rappel et Précision : c'est la précision moyenne une seule valeur reliant le rappel et la précision.

1.5.2 La collection TREC :

Le projet TREC (Text Retrieval Conference) est un programme international initié au début des années 90 par le NIST (National Institute of Standards and Technology) et du DARPA (Defense Advanced Reserach Projet Agency) [16]. Ce projet consiste en une série d'évaluations le moyen de mesurer l'efficacité de leurs systèmes de RI . Le projet TREC est constitué de différents éléments. Parmi ces éléments, les taches, les participants, la source d'information et enfin la structure et le principe de construction de la collection TREC. [17]

1.5.3 Autres campagnes :

On y trouve par exemple :

NCTR : Lancé en 1997, TREC sur des documents en Japonais (asiatiques).

Amaryllis : Lancé en 1996 et organisé par l'INIST. Amaryllis est la version française du projet TREC de 1996 à 1999. Il a pour objectif principal d'évaluer des logiciels de RI dans des corpus de texte en français.

CLEF (Cross Language Evaluation Forum) : Lancé en 2000, soutenu par l'UE. Il est pour l'évaluation de systèmes de recherche d'information multilingue.

INEX : Lancé en 2002 pour l'évaluation de systèmes de recherche d'information sur des documents XML

1.6 Conclusion :

Nous avons citer dans ce chapitre les différents principe de la RI classique particulièrement, introduit des notions de base(le besoin en information, la requête, le document et la pertinence..etc), ensuite décrit le processus de base de la RI. ainsi que définir les modèles de recherche d'information, Enfin, l'évaluation des systèmes de recherche d'information est traitée.

La recherche d'informations agrégée

2.1 Introduction :

À la lumière de l'évolution actuelle du Web ces dernières années, la recherche d'information (RI) traditionnelle ou bien classique basée pour la plupart sur des résultats de recherche est une liste des documents ordonnée correspondant aux besoins de l'utilisateur, mais actuellement, un seul document suffit. Toutes les informations pertinentes peuvent être trouvées dans différents documents, par exemple lorsque l'utilisateur recherche une ville à voyager, le résultat de cette recherche peut impliquer différent contenu (photos de la ville , actualités, météo, les hôtels, restaurants de la région, etc). Dans ce cas, le système doit fabriquer le résultat par la collections des informations auprès de plusieurs sources différentes afin de répondre le mieux possible à ses besoins, ce qui n'est pas le cas lors de la recherche d'informations traditionnelles. Ces limitations sont surmontées par un nouveau paradigme dans la recherche d'information multi-sources, ce qu'on appelle la recherche d'information agrégée, qui travaille sur des techniques d'agrégation des résultats de recherche. En d'autres termes, grouper les résultats de la recherche considère qu'il existe des moyens de rassembler les résultats pertinents et non redondants de différentes sources pour trouver l'information nécessaire pour la requête demandée par l'utilisateur. Nous présentons dans ce chapitre un aperçu sur les principaux travaux de recherche menés pour l'application du principe d'agrégation dans le contexte de la RI . En premier lieu, nous présentons les concepts de base, ainsi que le processus générique de la recherche d'information agrégée. Ensuite, nous montrons les problématiques liées à la RI agrégée. Enfin, nous citons les différentes approches et techniques de la recherche d'information

agrégée.

2.2 Qu'est ce que la recherche d'information agrégée ?

2.2.1 Définition :

Le paradigme de la RI agrégée a été défini pour la première fois dans l'atelier SIGIR'2008, qui considère la recherche agrégée comme suit « Aggregatedsearch is the task of searching and assembling information from a variety of sources and placing it in a single interface. » [5] La recherche agrégée permet de rechercher et d'assembler des informations à partir d'une variété de sources et en les plaçant dans une seule interface . Alors l'objectif ou bien la solution de la recherche agrégée est de fournir une vision plus riche et organisée des informations issues des différentes sources (images, vidéo) et de différentes granularités (passage de texte, entités, attributs, etc.), dans une seule interface .

2.2.2 Le processus générique de la recherche d'information agrégée :

Dans la recherche web, les moteurs de recherche fournissent certaines formes de la recherche agrégée, ils répondent à une requête d'utilisateur à travers la combinaison des résultats de différents moteurs de recherche, la recherche Web, la recherche vidéo, la recherche d'images. Alors la recherche agrégée peut être satisfaite sous différentes formes où chaque forme possède un processus propre à elle. D'après la définition de la RI agrégée devient claire, et son principe qui est assembler des informations provenant de diverses sources, en les plaçant dans une seule interface. A. Kopliku [17] a proposé un cadre général et un schéma conceptuel pour le processus de recherche d'information agrégée qui est compatible avec toute différente approches liée à la RI agrégée . la figure suivante présente un schéma conceptuel pour le processus de recherche agrégée Ce processus fonctionne comme suit : dans l'ordre, au début on exprime une requête qui peut être émise vers plusieurs sources d'information (moteur d'images, vidéo, etc), après chaque source récupère les informations pertinentes dans une ou plusieurs unités (nuggets) d'information et en fin les informations retournées doivent être assemblées, agrégées et organisées dans une seule interface pour fournir la réponse finale à l'utilisateur. On remarque que ce processus contient trois composants principaux pour la recherche agrégée qui sont, la répartition

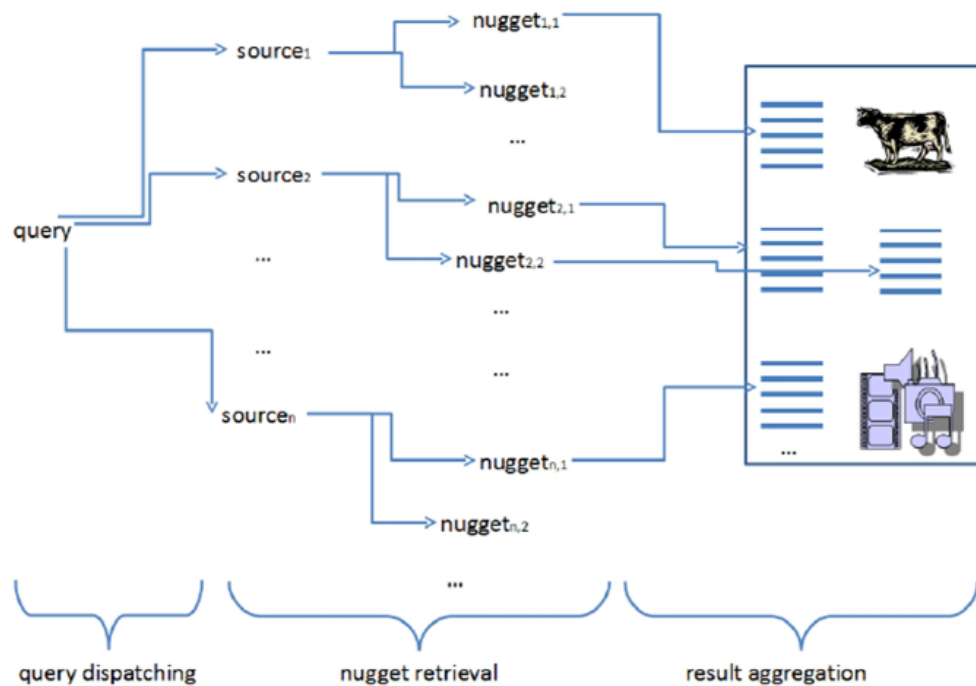


FIGURE 2.1 – le schéma conceptuel pour le processus de recherche agrégée

des requêtes (Query Dispatching (QD)), la recherche des unités d'information (Nuggets-Retrieval (NR)) et l'agrégation des résultats (ResultAgregation (RA)) :

La répartition des requêtes (Query Dispatching (QD)) :

Elle correspond à un traitement initial et une analyse de la requête. Cela implique différentes sous-tâches qui sont la sélection des sources et la décomposition des requêtes :

la sélection des sources : son objectif est de sélectionner les sources susceptibles de répondre à la requête. la sélection basée sur des technique qui sont l'identification des termes clés qui peuvent être utilisés pour sélectionner les sources, la présence de mots-clés spéciaux est souvent utile pour comprendre le type de réponse attendu par l'utilisateur, La recherche de graphe de Facebook reconnaît des termes tels que «j'aime», «amis» et «photos», l'identification des entités nommées et de leurs types peut améliorer le processus de la RI agrégée.

Décomposition des requêtes : Certaines requêtes peuvent être décomposées en deux ou plusieurs sous-requêtes. Par exemple, 'les musées et monuments d'Alger' peut être décomposée en 'les musées d'Alger' et 'les monuments d'Alger'. Si nous sommes incapables de trouver une page d'information sur les musées et les monuments d'Alger, nous pouvons joindre les résultats des deux sous-requêtes. Nous appelons ce type de requêtes : requêtes composées . [17]

la recherche des nuggets : La recherche des nuggets se situe entre la répartition des requêtes et l'agrégation des résultats. Pour une requête donnée qui correspond à une source précise. Nugget est un granule d'information qui répond à un besoin d'information ,la recherche des nuggets joue le rôle de collecter les informations pertinentes avec un scores de pertinences. La source renvoie un ou plusieurs nugget qui porte les informations les plus pertinentes pour cette requête, afin de placer les nuggets dans une interface unique qui contient l'agrégation des résultats .

Les résultats d'agrégations : Lors de la collecte de l'ensemble des nuggets d'informations pertinents, cela implique la récupération de ces nuggets de façon différentes, les assembler et l'agrégation des résultats de la recherche avant de les présenter dans l'interface pour que cela apparaisse à l'utilisateur. L'agrégation des résultats peut être faite selon la pertinence des nuggets sélectionnées, ou bien selon d'autres techniques que nous allons détailler ci-dessous.

2.2.3 Structure d'agrégat :

La recherche agrégée doit traiter l'organisation du contenu pertinent. Il n'est pas suffisant de collecter les informations mais il faut également les organiser pour la visualisation finale. La structure d'agrégat est définie comme toute l'information qui décrit le contenu, l'ordre de la visualisation, et les préférences dans la visualisation du document agrégé [18]. Les exemples suivants illustrent des structures partielles de certaines informations agrégées :

- 3 images, 2 vidéos.
- Une liste de 4 revues, 3 informations supplémentaires.
- Paragraphe A, paragraphe B, paragraphe C, avec l'ordre de visualisation : A, B,

C.

Dans la RI traditionnelle la réponse à une requête est une liste de documents. Dans la recherche agrégée il devrait être possible d'ajouter l'information de la structure d'agrégat à la requête. Quelqu'un pourrait demander des images et des vidéos seulement, ou des revues. Mais les études d'utilisateurs (Nielsen, 2003) indiquent qu'ils sont généralement paresseux. Ils n'emploient pas des options additionnelles. Néanmoins, ceci peut être très utile dans certains cas. Imaginez un service de voyage sur web qui recherche des hôtels et il est intéressé par des hôtels avec au moins trois photos, une carte, l'adresse, le numéro de téléphone, le nombre d'étoiles, les revues, etc. Ce type de recherche pourrait donc être utile dans ce cas.

2.3 les problématiques liées à la RI agrégée :

La RI agrégée pose aussi certaines problématiques qui doivent être mentionnées. A. Koplaku [17] cite quelques-unes : **Identification de type de réponse** : ce problème consiste à la question liée à l'identification des unités d'informations à renvoyer à l'utilisateur en réponse à sa requête. Certaines requêtes nécessitent une seule unité d'information, d'autres requêtes peuvent être répondues par plusieurs unités. La réponse à la question 'hauteur de la montagne de Djurdjura' ne nécessite qu'une seule unité, tandis que la réponse à la question 'restaurants traditionnels à Alger' nécessite de multiples unités. **Identifier les unités d'information les plus pertinentes** : les unités d'information qu'on peut récupérer lors de la RI agrégée ont de différentes granularités et de différents types. Il n'est pas triviale d'identifier les unités qui devraient être utilisées pour composer la réponse finale. Un document entier ou bien une unité d'information ? Quand est ce que l'utilisation du contenu multimédia (images, vidéos, etc.) est plus appropriée ? Quand devrions-nous utiliser les moteurs de recherche spécialisés (recherche d'images, recherche de vidéos, etc.) **Assembler les différentes unités d'information dans un document cohérent** : Dans la RI agrégée, plusieurs manières d'assembler les résultats de recherches sont possibles. Un résumé, deux images avec une définition, une table relationnelle, etc. Le choix de la meilleur agrégation selon les résultats de recherche disponibles est considérée comme l'un des objectifs de la RI agrégée. Quelle est la forme à laquelle le résultat final pourrait ressembler et évaluer la pertinence des résultats agrégés vis-à-vis

de la requête, sachant qu'il est impossible à priori de construire toutes les combinaisons possibles des résultats.

2.4 Les techniques d'agrégations et les approches d'agrégation

Le World Wide Web (WWW) est une source incroyable d'informations (documents), il peut être considéré comme un système d'information important distribué qui permet d'accéder à des données partagées par le biais du système de recherche d'information. L'utilisation des moteurs de recherche d'informations pour effectuer des recherches sur des requêtes simples, afin de retourner de nombreux documents, mais il est bien connu que dans le contexte de la recherche web, les utilisateurs accèdent à un nombre limité de documents dans l'espace de résultats, d'où il est difficile de trouver des informations pertinentes à cette recherche, il devient important de renvoyer des résultats plus divers sur ces pages pour obtenir une bonne satisfaction avec les informations disponibles sur le web sur un besoin d'utilisateur. La RI agrégée est l'une de ces techniques qui retournent des résultats de domaines variés (web, image, vidéo, news, etc.) et les présentent ensemble sur une page résultat. On cite quelques approches d'agrégation :

2.4.1 L'approche d'agrégation par clustering :

Cette approche a été proposée par Zeng et al, consiste à regrouper des documents après récupération des résultats de la recherche sous formes des clusters, afin de présenter un résumé de ces documents pour faciliter et orienter l'utilisateur à son groupe d'intérêts. L'utilité du regroupement dans la recherche d'information est de permettre la désambiguïsation et la facilité d'accès à l'information [19]. Parmi les systèmes de recherche d'information agrégée basé sur le clustering : **QCS (Query, Cluster, Summarize)** : qui réalise les tâches suivantes pour répondre à une requête donnée :

- Récupère les documents pertinents.
- Sépare les documents récupérés en groupes par sujet.
- Crée un résumé pour chaque cluster.

Yippy « **yippy.com** » : Clusty-yippy2 est un moteur de recherche qui interroge d'autres moteurs de recherche puis agrège, rassemble les résultats en groupes. Google news

est également un exemple intéressant d'agrégation par clustering. Il regroupe des nouvelles provenant de diverses sources. Les résultats ne sont pas une seule liste de nouvelles, mais un groupe de nouvelles. Chaque grappe concerne une histoire simple. Un bref résumé de l'un des éléments les plus représentatifs est donné en tant que titre de grappe. [20]

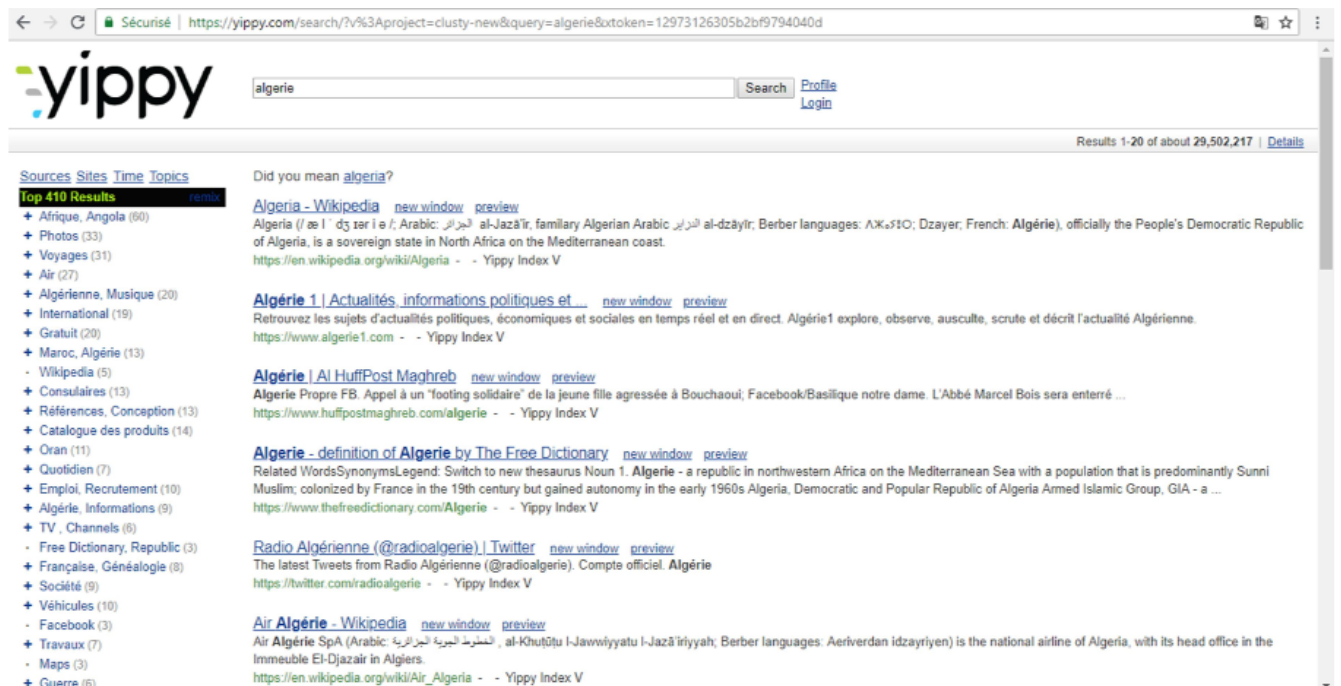


FIGURE 2.2 – exemple de yippy

2.4.2 L'approche d'agrégation par résumé multi-documents :

La mission générale de résumé multi-documents consiste à produire un résumé unique d'un ensemble de documents qui pourraient appartenir au même sujet. Les premiers systèmes de résumé automatique multi-documents ont été développés par Kathleen R. McKeown et Dragomir R. Radev dans les années 1990. WebInEssence est un système de résumé et de recommandation multi-document basée sur le web, qui utilise le résumé multi-documents comme technique d'agrégation afin d'aider les utilisateurs à trouver des informations utiles dans des documents spécifiques basés sur des profils utilisateurs personnels [20]. Cette technique est intéressante pour faire face aux problèmes de surcharge de l'information pour aider plus d'utilisateurs à trouver les informations qui répondent à leurs requêtes. Le résumé est le processus de choisir les informations les plus importantes du document.

Le processus de résumé multi-documents consiste en premier à choisir, évaluer, classer et assembler les unités de la langue (termes, phrases, paragraphes) selon leurs pertinences, en se basant sur des techniques statistiques. Ensuite on élimine les duplicités et les redondances, tout en gardant au mieux la cohérence de résumé.

2.4.3 La génération d'un document à partir de plusieurs documents :

La technique de génération de document cherche à créer ou automatiquement créer un document à partir des documents multiples, du même source ou des sources différentes [20]. Cette technique regroupe à la fois deux autres techniques qui sont l'agrégation par clustering et l'agrégation par résumé multi-documents. En réalité, elle s'agit de créer un document fictif à partir de plusieurs clusters issus des résultats par un moteur de recherche, ou chaque cluster correspond à des résumés de documents web retournés.

2.4.4 L'approche d'agrégation relationnelle :

La recherche agrégée relationnelle est une technique d'agrégation basée sur la relation entre les nuggets d'information récupérés, à la fin, elle fournit un résultat relationnel qui peut assembler des attributs associés à la requête, par exemple pour une requête "Algérie" que la recherche agrégée relationnelle fournit un résultat relationnel (attributs) nombre d'habitants , la superficie, etc Un résultat tabulaire agrégé de la forme "attribut / valeur" est construit pour chaque requête en trois étapes : La sélection des entités et des attributs pertinents pour la classe désignée par la requête, le filtrage des attributs récupérés et enfin, le tri des attributs pertinents. [20] RAS doit s'appuyer sur les avancées de nombreux domaines, notamment l'extraction d'informations, la recherche orientée entité, la recherche au niveau de l'objet, la recherche sémantique, la RI à partir des bases de données, La figure suivante est une exemples de recherche relationnelle agrégée sur Google Squared.

2.4.5 Les vues agrégées « Aggregated view » :

Les techniques que nous avons mentionné précédemment sont réalisées en fonction de générer une réponse (document) provenant de diverses sources . le traitement d'agrégation et la diversification des résultats de recherche avec différentes standards moteurs verti-

The image shows a screenshot of the Google Squared search interface. At the top, the Google Squared logo is on the left, followed by a search bar containing the text 'US presidents'. To the right of the search bar are two buttons: 'Square it' and 'Add to this Square'. Below the search bar, the results are displayed in a table format with the following columns: Item Name, Image, Description, Date Of Birth, and Preceded By. Two results are visible: George Washington and Abraham Lincoln.



Item Name	Image	Description	Date Of Birth	Preceded By
George Washington		George Washington was born on February 22, 1732 [O.S. February 11, 1731] the first child, of Augustine Washington and his second wife, Mary Ball Washington, ...	February 22, 1732	Charles Denison
Abraham Lincoln		Abraham Lincoln (February 12, 1809 – April 15, 1865) was the 16th President of the	February 12, 1809	James Buchanan

FIGURE 2.3 – exemples de recherche relationnelle agrégée sur Google Squared

caux , un moteur vertical est un moteur de recherche qui peut être un moteur d'images, vidéos, ...etc. l'approche des vues agrégées permet de fusionner et d'interroger les différents résultats agrégés de moteurs verticaux dans une seule page de résultats. Cette technique est utilisée dans la recherche agrégée Web. A la fin, le résultat de la recherche sera présenté comme une vues agrégée . J. Arguello et al [21] et M. Lalmas et al [22] ont décrit deux types de représentation des vues agrégées : la "blended view" et la "non-blende dview"

blended view« mixte » :

Google universal search et beaucoup d'autres moteurs de recherche ont appliqué la conception mixte et présentent le design mixte aux utilisateurs, dans la présentation des résultats de recherche dans un design mixte, ces derniers sont regroupées, représentées de manière hétérogènes malgré ils sont récupérés de différentes sources (ou moteurs de recherche verticaux), comme Google images ,Google vidéos , Google map .. etc.

Environ 74 100 000 résultats (0,58 secondes)

Page d'accueil | UNICEF
<https://www.unicef.org/fr> ▼
 Découvrez les actions menées par l'UNICEF pour défendre les droits des enfants et protéger la vie de chacun d'entre eux, chaque jour. Child protection icon.

À propos de l'UNICEF
 L'UNICEF travaille dans 190 pays et territoires pour sauver des ...
[Autres résultats sur unicef.org >](#)


Fonds des Nations unies pour l'enfance — Wikipédia
https://fr.wikipedia.org/wiki/Fonds_des_Nations_unies_pour_l%27enfance ▼
 Cet article ne cite pas suffisamment ses sources (décembre 2015). Si vous disposez ... L'Unicef a reçu le prix Nobel de la paix le 26 octobre 1965.
[Stratégies de ...](#) · [Missions](#)

À la une

L'UNICEF accepte désormais les dons en crypto-monnaies
 Clubic · Il y a 3 jours

UNICEF : ouverture d'une campagne de crypto-dons
 JournalduCoin.com · Il y a 3 jours


Beni : 155 enfants sont orphelins ou séparés de leurs parents suite à


Fonds des Nations unies pour l'enfance 


Le Fonds des Nations unies pour l'enfance est une agence de l'Organisation des Nations unies consacrée à l'amélioration et à la promotion de la condition des enfants. [Wikipédia](#)


Siège social : New York, État de New York, États-Unis
Type d'organisation : Organisme public international
Organisation parente : CÉSNU
Président : Tore Hattrem
Création : 11 décembre 1946, New York, État de New York, États-Unis
Fondateurs : Assemblée générale des Nations unies, Ludwik Rajchman


Recherches associées [Voir d'autres éléments \(plus de 10\)](#)


 Organisat...
des Nations
u...


 Organisat...
des Nations
u...


 Programme
alimentaire
mondial


 Save the
Children


 Médecins
sans
frontières

[Clause de non-responsabilité](#) [Commentaires](#)
[Revenir à cette fiche info](#)

FIGURE 2.4 – exemple Google :design blended

Non-Blended « non mixte » :

D'autres moteurs de recherche comme Yahoo Alpha5 et ASK3D6 utilisent une "vue non-blended" ou non mixte. La conception de non-blended vue qui présente les résultats de chaque recherche verticale séparément dans les panneaux de la page de résultat de recherche, ces différents panneaux sont regroupés selon le type images ,vidéo. . . et s'affichent comme des résultats. Le placement des différents groupes est prédéfini. La figure 3 montre le résultat de la recherche sur YahooAlpha qui est un exemple d'une telle conception.

Search results for "world cup 2018" on Yahoo. The page displays a search bar at the top, followed by navigation tabs (Web, Images, Video, News, More, Anytime). Below the search bar, there are several vertical panels:

- FIFA World Cup - Schedule & Results:** Shows match results for June 22 and June 24. The June 22 results include Belgium 5 vs Tunisia 2, Germany 0 vs Sweden 0, South Korea 1 vs Mexico 2, and a Final match.
- FIFA Live Standings:** A table showing the standings for Group A.
- World Cup - News:** A section with news snippets, including "Germany vs Sweden, World Cup 2018: live score and latest updates" and "World Cup 2018 fixtures: full match results and complete schedule".
- FIFA World Cup:** Overview information for the 2018 tournament, including the host country (Russia), network (Fox Sports, Telemundo), teams (32 qualified teams), official mascot (Zabivka), official song (Live It Up), and the 2014 champion (Germany).
- Past World Cups:** A section displaying logos for previous World Cup tournaments from 2014 to 1998.

FIGURE 2.5 – exemple de alphaYahoo(design non-blended)

2.5 Conclusion :

Dans ce chapitre nous avons présenté un bref aperçu sur la RI agrégée avec le processus générique de la recherche d'information agrégée. Nous avons aussi montré les problématiques liées à la RI agrégée. Et enfin, nous avons cité les différentes approches et techniques de la recherche d'information agrégée. Dans le prochain chapitre, nous présenterons l'e-tourisme et l'application de la recherche d'information agrégée sur les brochures touristiques.

Brochure touristique

3.1 Introduction :

Dans le siècle actuel, la technologie a envahi tous les aspects de la vie humaine dans divers secteurs, le tourisme parmi les secteurs les plus touchés. D'après les statistiques faites par OMT (l'Organisation Mondiale du Tourisme) , le nombre de touristes a augmenté de 25 millions de touristes internationaux en 1950 à 1087 millions en 2013, car la croissance du tourisme dans le monde a permis à ce secteur de devenir important dans l'économie nationale . En raison de la gestion traditionnelle du tourisme à travers les agences touristiques qui ont causé plusieurs problèmes, et le développement technologique (Internet), les internautes se tournent vers les innovations technologiques pour subvenir à leurs besoins, ce qui a conduit à l'émergence du e-tourisme qui regroupe un ensemble des activités touristiques (réservation, hébergement, restauration, . . . etc.) liées à l'e-commerce grâce à l'internet. A cause de l'importance de la publicité et du marketing à travers les guides et les brochures touristiques et à la lumière du développement technologique et de l'e-tourisme, les touristes peuvent facilement accéder aux informations sur les lieux à visiter tels que les hôtels, les restaurants...etc. Grâce à la génération automatique des brochures touristiques afin de gagner du temps pour les touristes et d'économiser l'argent dépensé sur les moyens écrits pour les agences touristiques. Dans ce troisième chapitre nous allons traiter la notion du e-tourisme, la brochure touristique, ses objectifs et ses intérêts, présenter par la suite une étude d'une brochure touristique afin d'identifier son contenu.

3.2 L'e-tourisme :

3.2.1 Du Web 1.0 au Web 3.0 :

Le World Wide Web (WWW) est démocratisé en 1989 par Tim Berners-Lee, qui a créé comme "un système hypertexte public fonctionnant sur Internet qui permet de consulter, avec un navigateur, des pages accessibles sur des sites."

- Le web 1.0 (web traditionnel) est appelé aussi le web statique, il inclut les pages statistiques des sites web, les propriétaires de ces sites font toute la gestion sur le site. . . etc, ces sites nécessitent peu d'intervention des utilisateurs, ils ne peuvent que consulter le site. [23] Dans le cas de l'e-commerce (achat en ligne) les utilisateurs sont des consommateurs des produits.
- Le web 2.0 (le web social) parce qu'il est plus orienté partage de contenu (textes, images, vidéos, . . . etc.) et l'échanges des informations et aussi permet aux utilisateurs l'interaction et la conversation entre eux. Dans le web 2.0 l'utilisateur n'est plus un simple consommateur, il devient un propriétaire de son propre contenu. Dans cette optique sont apparus les outils suivants : les réseaux sociaux, plateformes interactives (lire, écrire, et partage), sites dynamiques, blogs, wikis, . . . etc
- Le web 3.0 (web sémantique) Ce concept émerge depuis les années 2008. C'est le Web en temps réel. Les systèmes sont interopérables. Une véritable société numérique se met en place au sein de laquelle humains et agents intelligents collaborent pour générer des connaissances utilisables par les humains et les machines. Tout le monde devient, à la fois, consommateur et producteur. [24]

3.2.2 Le Web au service du tourisme en ligne :

L'apparition du Web 2.0, les changements et les développements dans les innovations technologiques ont affecté la culture de consommation des gens et ils les ont complètement changés. Le Web au service du tourisme en ligne : Les personnes s'orientent désormais vers le Web afin de répondre à leurs besoins, ils ont trouvé que le web est un moyen simple de couvrir leurs besoins dans tous les domaines. Le secteur touristique est parmi les secteurs les plus dominants du e-commerce (commerce électronique) engendrés par la révolution du web ,grâce aux sites. web qui permettent aux utilisateurs d'intervenir en temps réel en utilisant des fonctionnalités qui consistent de collecter toute les informations liées au

tourisme, comme la réservation en ligne des hôtel, commander des repas, des transports aériens, automobiles, des locations de véhicules, des séjours en camping, des activités de loisirs, ou des visites de musées . . . etc, et mettre à leur disposition plus d'informations sur une destination préférée de voyage via des outils d'internet . C'est dans ce contexte que l'e-tourisme a vu le jour.

3.2.3 Qu'est ce que le e-tourisme ?

L'e-tourisme ou le tourisme électronique c'est la promotion, la commercialisation, la vente de services touristiques en ligne. Les services touristiques qui concernent toute les activités liées au tourisme pendant tout le voyage (hébergement, restauration, loisir, transports aériens, automobiles. . . etc). Autrement, le e-tourisme offre les services touristiques en ligne pour préparer, organiser et réserver ses voyages, la visibilité de la destination via internet, et garantir certain objectif comme trouver de bonnes prestations et sélectionner la meilleure destination en se basant sur des photos, des images, des vidéos. . . etc, ainsi d'avoir des prix avantageux et comparables. Parmi les outils qu'un touriste utilise pour s'informer sur sa destination on trouve le Web et les brochures qui sont les moyennes le plus utilisés. Dans l'e-tourisme les brochures sont aussi électroniques et automatique, qui peuvent satisfaire et répondre aux besoins des touristonaute,touristonaute veut dire Les touristes qui utilisent le net pour se renseigner sur une destination.

3.3 Les brochures touristiques :

3.3.1 Définition :

La brochure touristique est une brochure avec un contenu touristique qui contient toutes les informations importantes sur un lieu bien précis, aussi elle contient des divers types d'informations comme les images, vidéos, carte géographique, météo etc.ces informations sont sélectionnées et structurées afin de donner un aperçu aux touristes.

3.3.2 Les objectifs et les intérêts d'une brochure touristique :

La brochure touristique est un outil de publicité et marketing qui aide les touristes à planifier des voyages et leurs donne une vue d'ensemble sur le transport aérien et terrestre,

la localisation géographique électronique qui contient tous les détails sur la zone de destination ainsi que l'hébergement et la restauration, les lieux touristiques et aussi fournit des suggestions des plans des sorties, concerts, événements divers, visites aux musées, . . . etc. En effet, beaucoup des voyageurs recherchent de l'information avant et durant le voyage, selon les statistiques représentées dans la figure ci-dessous, des études révèlent qu'une fois le domicile quitté, les brochures deviennent l'outil de planification par excellence pour prendre une décision. Les objectifs d'une brochure touristique ne diffèrent pas de ceux de

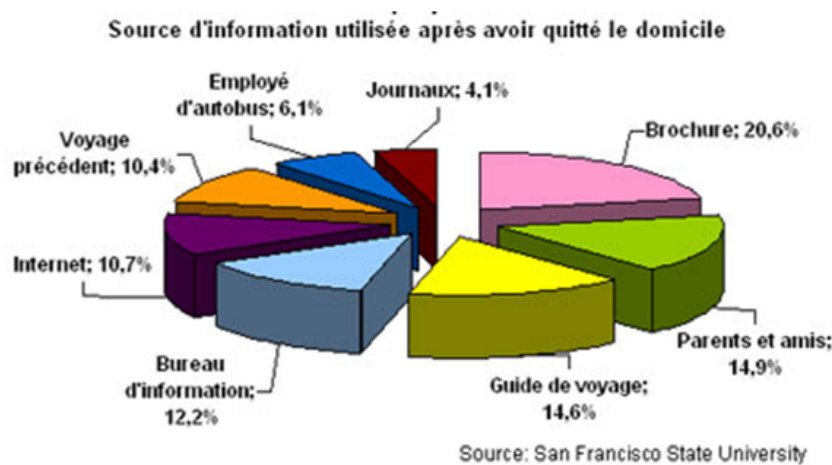


FIGURE 3.1 – les sources des information utilisées après avoir quitté le domicile

la publicité pour d'autres produits. [25] a résumé les objectifs sous-jacents de la publicité en trois mots : «informer, de persuader et de rappeler», qui sont en ligne avec le principe AIDA utilisé dans le marketing : "attirer l'Attention, faire naître l'Intérêt, susciter le Désir et entraîner l'Acte d'achat» [25].

- **Inform** : la brochure touristique son rôle principal est d'informer les touristes et fournit les informations adéquates lors de la lecture des ces différentes pages. Les touristes n'avaient aucune idée d'où la brochure était destiné à donner une couverture publique et à leur informer comme exemple les différents hôtels existes, les restaurants avec les horaires d'ouverture, les événements, la météo, . . . etc. autrement dit, tous les différentes activités à pratiquer dans la région, globalement fournir les informations pratiques dont le lecteur aura besoin.
- **Persuader** : parmi les objectifs attendus par la brochure touristique est de convaincre et d'encourager les touristes à visiter le lieu de destination ou de laisser une mer-

veilleuse impression de rendre visite, surtout le côté visuel de fournir de belles images et le tourisme le plus possible dans cette région et toutes les installations pour atteindre la satisfaction des touristes.

- **Rappeler** : est utilisé pour rappeler aux touristes existants un les places a visiter ou un service (hébergement,...) et rappeler les meilleurs les endroits et bonne réception, cet objectif tente de garder et gagner la confiance du visiteur.

3.3.3 Contenu et structure d'une brochure touristique :

Après la consultation de plusieurs brochures touristiques, nous avons sélectionné deux brochure touristique riches en informations tant textuels que visuelles, combinant les informations nécessaires et bien précises et fournit une façon de présenter le contenu très organisé et cohérent avec un design qui attire l'attention. Contenu et structure d'une brochure touristique : Dans ce suit on analysera le contenu textuels et le contenu visuels de la brochure «Amsterdam »[26] et la brochure « London »[27].

Pourquoi la brochure « Amsterdam » ? :

nous avons évalué les deux brochures touristiques à plusieurs égards, en choisissant la brochure touristique sélectionné est distinguée par leur contenu intégré et complète en fournissant des informations précieuses et utiles et une communication textuelle et visuelle qui contribue à l'interaction des touristes, la structure de présenter leur contenu est bien organisé et de ce design agréable qui joue un rôle important dans l'attention des touristonautes . On analysera dans ce qui suit le contenu textuel et le contenu visuel de la brochure Amsterdam [26] Pour plus d'objectivité dans notre étude et analyse, nous devons prêter l'importance au contenu textuel ainsi qu'au contenu visuel. Cette section est divisée en l'analyse de contenu textuel et l'analyse de contenu visuel de contenu de la brochure Amsterdam.

La brochure « Amsterdam » :

Amsterdam est la commune la plus peuplée et la capitale du Royaume des Pays-Bas bien que le gouvernement ainsi que la plupart des institutions du pays siègent à la Haye. La ville est située en Hollande-Septentrionale .

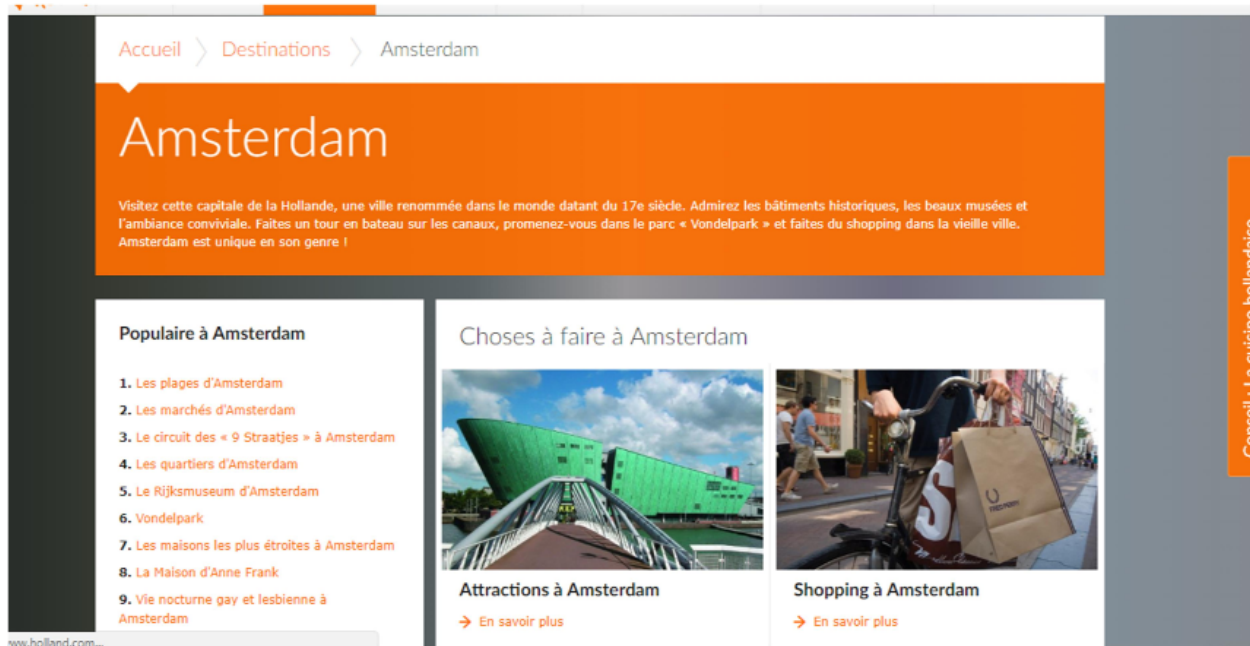


FIGURE 3.2 – brochure Amsterdam

Contenu textuel : On constate que le contenu a été repartitionné en 27 sections, chaque section contient une collection de pages et de documents liés au titre de la section et avec un sommaire qui contient les éléments essentiels et populaires pour faciliter la recherche et la navigation dans la brochure. Nous mentionnerons certains des sections :

- **Top choses à voir à Amsterdam :** mentionner dans un petit sommaire les 10 tops places populaires et préférées par les touristes.
- **Choses à faire à Amsterdam, À ne pas manquer à Amsterdam, Expérience unique à Amsterdam :** sont des sections qui représentent des attractions de différents types et les lieux touristiques qui distinguent la zone qui reste dans la mémoire avec une description de chaque zone et son adresse... etc.
- **Plan d'Amsterdam :** la carte géographique sous Google map avec les positions de toutes les attractions touristiques.
- **Découvrez Amsterdam :** représentation et description, histoire d'Amsterdam.
- **Manifestations en Hollande :** calendrier des festivals avec toutes ces informations.
- **Se déplacer à Amsterdam :** tout ce qui concerne le transport avec les tarifs.

- **Les quartiers d'Amsterdam** : on trouve la carte géographique de chaque quartier et leurs propres hôtels etc.
- **Hotels recommandés à Amsterdam** : c'est l'hébergement, il existe plusieurs hôtels avec la possibilité de réservations.
- **Réaurants à Amsterdam** : on y trouve les restaurants et les salons de thé les plus visités,
- **Sortir à Amsterdam** : des suggestions des plans des sorties, concerts, événements divers.

Contenu visuel : Les images sont un élément clé dans la brochure, la sélection des images est faite soigneusement, premièrement, on n'utilise que des photos qui représentent la destination, ces photos modérée et ayant une bonne résolution attirent l'œil de l'utilisateur (future client) et contient en quelque sorte une petite histoire qu'on veut raconter sur le lieu en question. Ensuite, ces photos sont vives, ils contiennent des personnes souriant, amusant mieux que des photos mortes (une chambre d'hôtel vide ou une plage). Les couleurs sont en harmonie avec les images, ainsi que leur utilisation est indicative, chaque lieu propose un ton différent illustré par la couleur qui convient. De même la couleur principale utilisée qui est l'orange est unique dans toutes les sections de la brochure, qui caractérise le drapeau des Pays-Bas (Holland) et puisque trop de couleurs cela pourrait être distrayant et tape-à-l'œil. Dans la figure représente la section de restaurants à Amsterdam on trouve les photos tout ce qui concerne la nourriture, la boisson, la nourriture traditionnelle, les restaurants et les cafés . . . etc.

3.3.4 Synthèse de l'étude de la brochure :

La synthèse qui constitue un résumé de notre étude, grâce à l'analyse de la brochure touristique précédente, nous distinguons clairement l'émergence de plusieurs points importants, en termes de contenu, la brochure est divisé en trois catégories sont contenu de base, contenu spécifique et caractéristique de la région, et contenu supplémentaire.

Contenu de base : les éléments principaux nécessaires qui doivent être disponibles et fournies par chaque brochure touristique, parmi ces éléments on a hébergements, les restaurants, transport et ses moyens, carte géographique . . . etc.

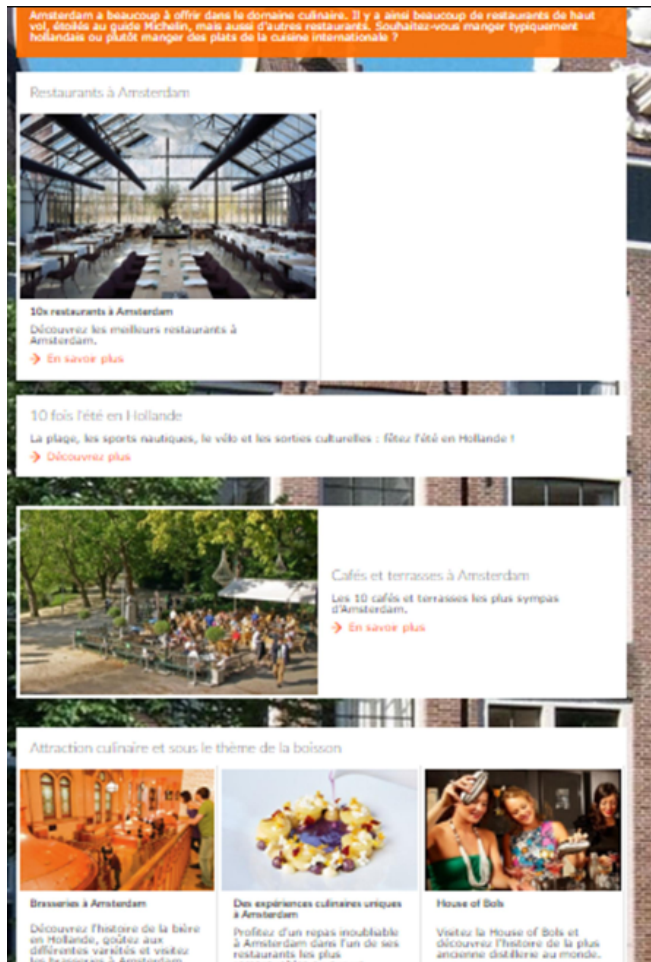


FIGURE 3.3 – la section de restaurants à Amsterdam

Contenu spécifique et caractéristique de la région : les éléments qui caractérisent la région, Qui définit toutes les dimensions de la région et ses principales caractéristiques les plus distinctives, en particulier touristiques comme les musées, marché, shopping, . . . etc.

Contenu supplémentaire : ces éléments est non essentielles mais qui sont des luxes et des améliorations et clarifier l'utilisateur plus facilement pour trouver son besoins, parmi ces éléments on a météo, résumé, artisanat, festivals, . . . etc.

3.3.5 Limitations des brochures touristiques classiques :

La brochure touristique joue un rôle majeur dans la détermination et la décision de la destination des touristes, car ils préfèrent avoir une brochure de contenu complet et informatif qui peut leur être utile, mais la brochure que nous avons analysée malgré la

disponibilité des informations nécessaires et riche mais reste limité, nous citerons certaines des limites les plus communes auxquelles nous sommes confrontés dans la brochure classique.

La disponibilité : les brochures classique (hors ligne) limité à un certain petit nombre des régions.

La mise à jour : le contenu des brochures est statique, Ce qui signifie que ne change qu'en élaborant une nouvelle brochure après une année ou plus.

3.4 Conclusion :

Comme nous l'avons vu précédemment la recherche agrégée est une technique qui nous permet de rassembler des données de différents types et de différentes sources. La brochure est un cas pratique de la recherche agrégée car l'information touristique est une information diversifiée et variées (texte, images, ... etc.), où dépend de sa conception sur la sélection les attribues de la brochure de différents types et sources, afin regrouper et structurer les données dans une interface de la brochure. dans le contexte de la recherche d'information agrégée, on s'est proposé de développer un système de génération automatique de brochures touristiques en appliquant le paradigme de recherche d'information agrégée, système correspond a un mini moteur de recherche utilisant la recherche d'information agrégée pour répondre a une requête de brochure touristique. Dans ce chapitre on a traité la notion et l'émergence du e-tourisme, nous avons présenté, la brochure touristique, ses objectifs et ses intérêts, après on a fait l'analyse et l'étude de la brochure touristique Amsterdam et déduire la synthèse le contenu nécessaire pour concevoir un générateur automatique de brochures touristiques. Grace a ces analyses, nous comprenons l'importance de la RI agrégée pour développer un système de génération automatique de brochures touristiques qui pouvant rassembler et agrégée les données et visualiser la brochure à partir des différents sources. En effet la génération automatique de brochures touristique est un très bon exemple d'application de la RI agrégée pour au moins deux raison :

- Le contenu d'une brochure est un contenu hétérogènes et diversifié (texte, image, carte, données structurés (liste d'hôtels), météo, ...). La RI agrégée essaye justement d'agréger des informations hétérogènes dans le même espace de résultats

— Pour arriver a produire une brochure touristique nous avons besoin de puiser de l'information à partir de plusieurs sources. La multi-source est justement à la base de la RI agrégée.

— L'agrégation du contenu est au cœur des tous systèmes de RI agrégée, il se trouve aussi que cette notion est un élément de base dans le cas d'une brochure touristique

Dans le suivant chapitre nous proposons et présentons la conception de notre solution et l'organisation de la brochure, la modélisation du système.

Conception et modélisation

4.1 Introduction

Nous avons montré dans le chapitre précédent que la génération automatique de brochures touristiques peut constituer un bon exemple d'application du paradigme de la RI agrégée. En effet pour pouvoir produire automatiquement une brochure nous avons besoin de :

- Chercher et sélectionner de l'information variée et hétérogènes (description, images, cartes, restaurant, météo ,) à partir de plusieurs sources (le multi-sources)
- D'agrèger toutes ces information hétérogènes dans un même espace de réponse bien organisé et cohérent

Comme on l'a mentionné dans le chapitre deux consacré à la RI agrégée, trois problématiques majeurs existent en RI agrégée qui sont :

- la problématique de représentation des sources (comment représenter mes sources ?).
- la problématique de Sélection des sources (quelles sources utiliser pour une requête donnée).
- Enfin la problématique de fusion ou agrégation des résultats.

Pour notre travail on va plus s'intéresser à la dernière (fusion ou agrégation des résultats).

4.2 La conception de notre brochure :

Le contenu de la brochure touristique doit être complet, structuré, organisé et très convaincant pour l'utilisateur, il doit également être informatif, y compris les éléments

importants qui satisfont les besoins, ce rôle étant également contenu textuel et visuel. Nous avons pour notre cas opté pour les contenus suivant d'après le chapitre précédent.

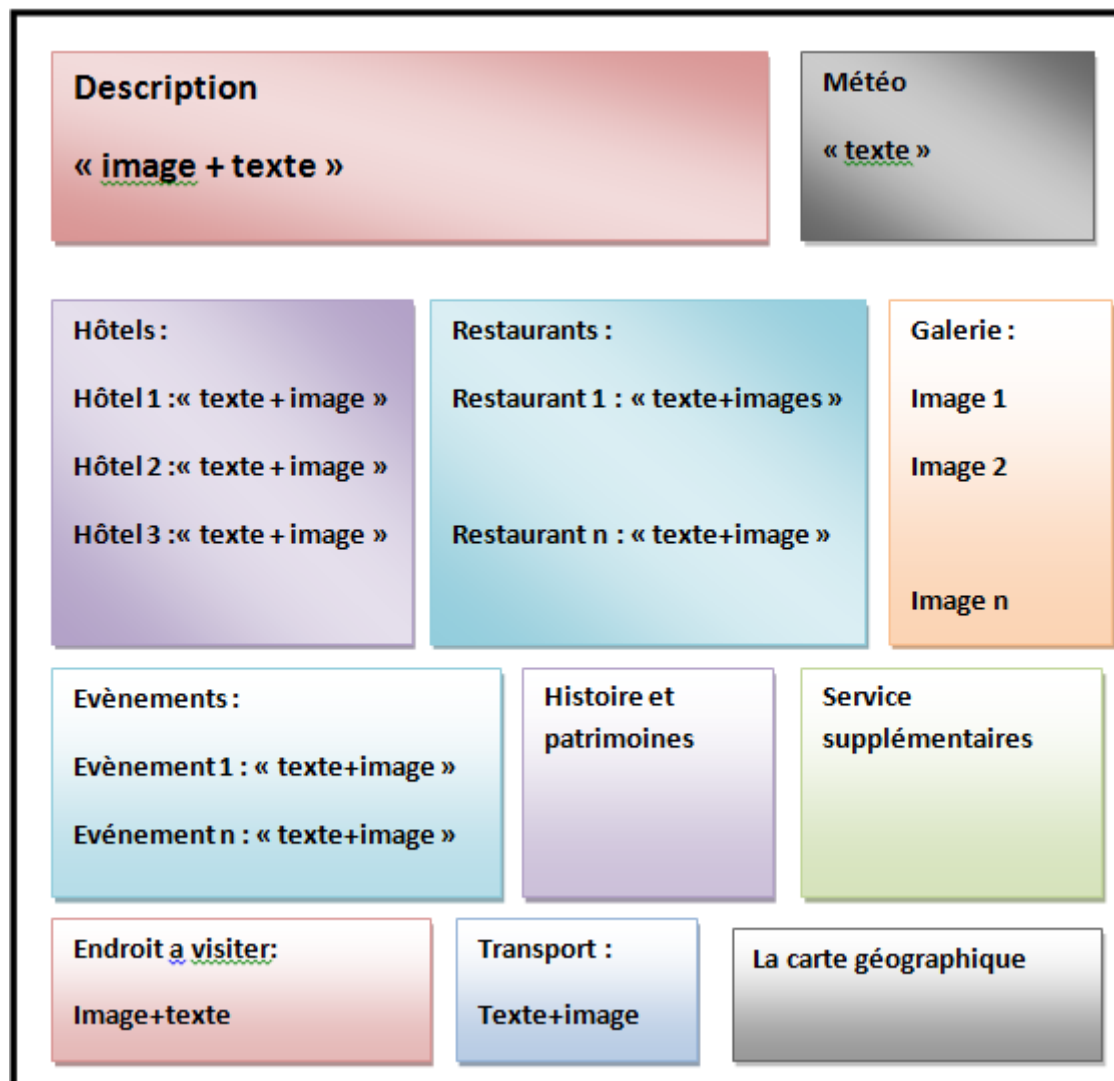


FIGURE 4.1 – le contenu visuel et textuel des notre brochure

4.3 Présentation générale du système :

Etant un cas d'application de la Recherche Agrégée, nous nous sommes basés sur le processus générique de la RI agrégée proposée par Arlind Koplaku pour développer notre système. Ce processus se compose de trois phases :

- la phase de dispatching de la requête
- la phase de recherche de nuggets (les items résultats)

— la phase d'agrégation des résultats

Nous allons donc adopter ce processus et développer chaque étape. Mais avant de développer chacune de ces étapes, nous avons besoin tout d'abord de choisir le contenu des brochures à générer et aussi surtout de choisir les sources d'information à utiliser pour chacun des contenus.

4.3.1 Le processus de génération de la brochure :

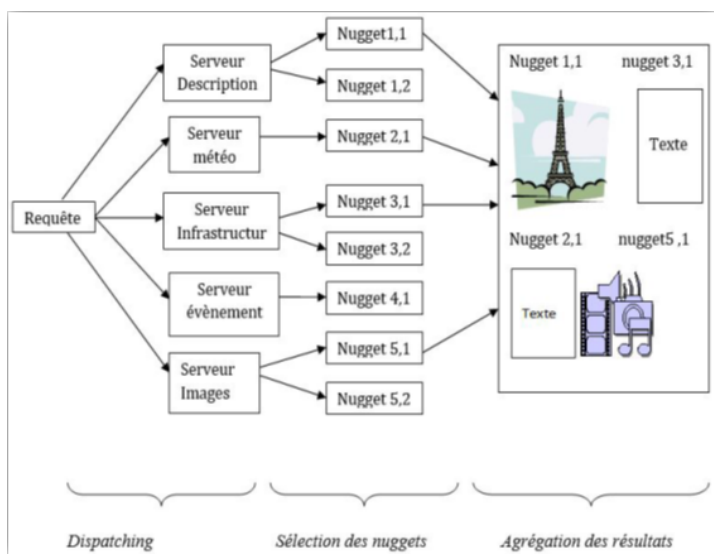


FIGURE 4.2 – Le processus de génération de notre brochure

4.3.2 Les sources d'informations utilisées :

Dans notre travail, la recherche se fait sur le web qui représente la source de donnée la plus utilisée. Nous allons s'intéresser à quelques moteurs de recherche qui offrent l'information souhaitée dans le domaine touristique tel que les descriptions des lieux, les images, les cartes géographiques, la météo, les événements... etc. On a identifié la(es) source(s) de chacun de ces nuggets, le nugget est un granule d'information répond aux besoins un besoin d'information, chaque nugget possède une ou plusieurs sources, les résultats des requêtes vers ces sources seront à la fin agrégés pour construire un résultat agrégée.

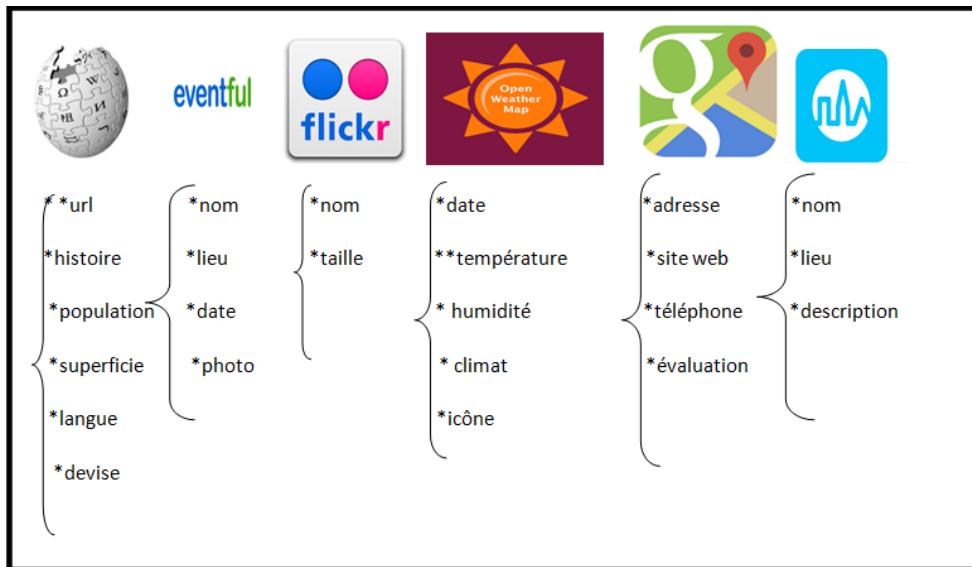


FIGURE 4.3 – identification des sources

4.3.3 Dispatching

Cette étape dans notre système consiste en l'interprétation de la requête saisie par l'utilisateur sous forme de texte qui exprime le nom d'une région, ensuite la requête sera convertie vers des coordonnées (la latitude et la longitude) correspondants à la zone géographique de la région saisie. Ensuite les coordonnées sont envoyées vers plusieurs sources sous formes des sous requête, chaque sous requête est traitée dans une source et renvoie le résultat de la recherche correspond à une type d'information selon le serveur (source). Nous prenons comme exemple la ville «Taghit», que l'utilisateur veut visiter et prendre une vue global sur le tourisme dans cette, à partir de notre application cet utilisateur saisit la requête «Taghit », ensuite le mot saisi sera converti, la requête sera par la suite émises aux différentes sources qui alimentent notre application.

4.3.4 Sélection des nuggests :

Nous interrogeons les différentes sources en utilisant des API (Application Programming Interface). Ces API qui sont conçues par les moteurs de recherche pour faciliter la tâche du développement, et pour économiser plus de temps et d'efforts, et offrent une logique de programmation adéquate. Dans notre travail, nous aurons besoin de plusieurs nuggests (unités d'information), tel que les images, les infrastructures, la météo... etc.

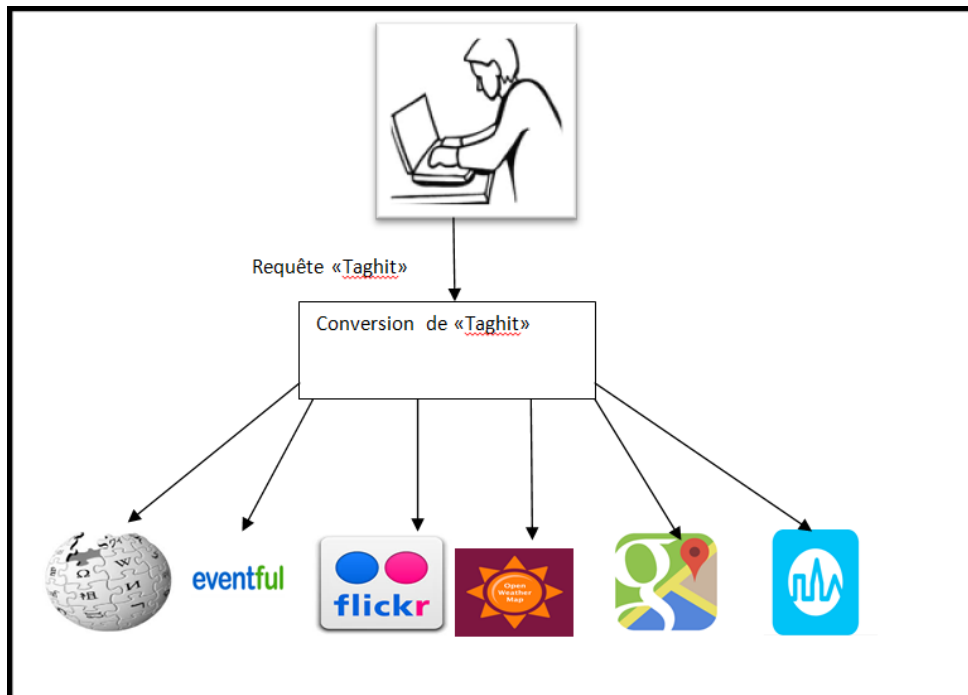


FIGURE 4.4 – le dispatching de la requête «Taghit»

Chaque nugget contient des informations selon sa granularité et ils sont récupérés de plusieurs sources, donc il est nécessaire de faire en même temps plusieurs requêtes vers plusieurs moteurs de recherche, chaque sous requête aura comme réponse un ensemble de nuggets qui seront utilisés dans l'étape d'agrégation pour construire notre brochure. chaque sources d'information va restituer non pas un mais plusieurs résultats (exemple : des centaines de résultats texte, des centaines d'images, des centaines fiches d'hôtels,,,,). Nous devons donc sélectionner parmi les résultats de chaque source ceux qui devront figurer sur la brochure. Les critères de sélection des résultats sont multiples :

- Tout d'abord : la pertinence. Nous devront choisir de chaque source les résultats les plus pertinents. Puisque nous nous basons sur des API fournées pas ces sources et donc sur des algorithmes de recherche qui sont supposés donner les meilleures résultats en premier lieu,
- Critères géographique : nous devons choisir par exemple parmi les hôtels ou restaurant retournées ceux qui sont les plus proches géographiquement du lieu objet de la requête.
- Le critère d'espace : une brochure est un document concis qui a un espace limité. Nous devons donc nous limiter a seulement certains éléments de chaque sources

(exemple : 10 hôtels, 10 restaurant, 10 premières images, ...).

La source de description et la source météo, renvoient un seul résultat qui est pris automatiquement. Pour ces deux sources il n'y a pas vraiment de sélection. Mais pour les sources de google place on sélection les nuggets selon le critère géographique "le plus proche" au lieu objet du requeté, après on choisir le n premier selon le critère d'espace. Pour la source des évènements et la source des images on sélection les n premiers résultats (critère de pertinence et d'espace).

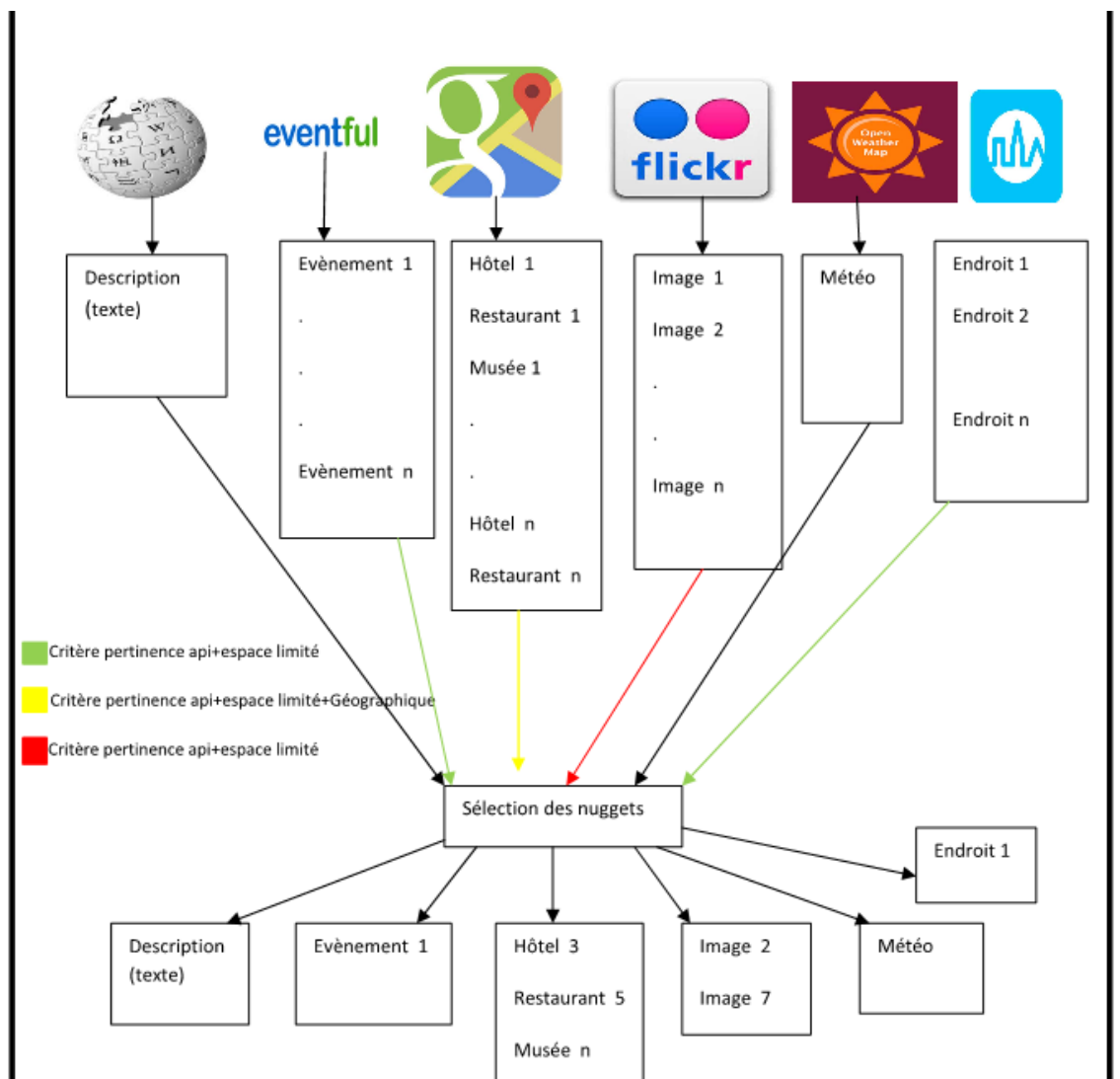


FIGURE 4.5 – la sélection des nuggets

4.3.5 L'agrégation des résultats :

Là phase d'agrégation des résultats consiste à l'agrégation des nuggets qui ont été sélectionnés dans la phase précédente à partir de différents sources dans seul interface. L'objectif de l'agrégation des résultats est d'organiser les données collectées d'une manière facile qui permet à l'utilisateur d'avoir un espace de réponse organisé, complet et bien cohérent. La brochure touristique contient plusieurs parties ou panneaux (description, hôtel, transports. . . etc). Parmi les approches d'agrégation utilisées en RI agrégée, l'agrégation via les vues agrégée de type non blended semble l'être la plus adaptée à notre cas. En effet cette dernière devise l'espace des résultats en plusieurs panneaux, chaque panneau est destiné a recevoir un type d'information particulier. De la même notre brochure est constituée de plusieurs zones et chaque zone traite un aspect particulier.

Nous pouvons aussi combinée dans cette approche l'agrégation relationnelle pour la partie hôtel, restaurant et même évènement. Ces trois zones peuvent être affichées sous un format tabulaire la figure suivante montre la structure de notre brochure qui est ensemble des panel :

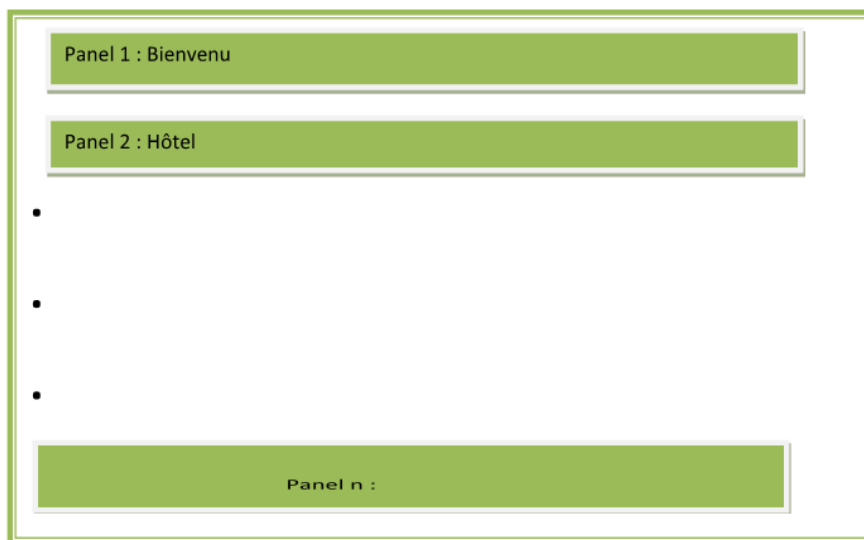


FIGURE 4.6 – la structure de la brochure

Agrégation Non Blended :

Dans les vues agrégées non-Blended (non-mixte) les résultats sont regroupés par rapport à leur type (images, news, vidéos, etc) et représenté les résultats de chaque verticale

séparément dans le panneau de la page de résultats de recherche, ces différents panneaux sont regroupés les résultats (voir le chapitre ‘la recherche agrégée’). Par exemple dans la figure ci-dessus représente le panel galerie :



FIGURE 4.7 – la structure de la galerie d’images

4.4 La modélisation UML (Unified Modeling Language) de système :

UML est une notation graphique conçue pour représenter, spécifier, construire et documenter les systèmes logiciels. Ses deux principaux objectifs sont la modélisation de systèmes utilisant les techniques orientées objet, depuis la conception jusqu’à la maintenance, et la création d’un langage abstrait compréhensible par l’homme et interprétable par les machines. la figure ci-dessus montre les différents types des diagrammes Uml (statique, dynamique et fonctionnel). Dans la suite, nous allons modéliser notre application en utilisant un diagramme de chaque type.

4.4.1 Les objectifs d’UML :

UML est un langage formel et normalisé qui permet durant la phase de conception :

- Un grain de précision.
- Un gage de stabilité.

Le langage UML est un support de communication performant :

- Il encadre l’analyse.
- Il facilite la compréhension de représentation abstraite complexe.
- Son caractère polyvalent et sa souplesse en font un langage universel.[28]

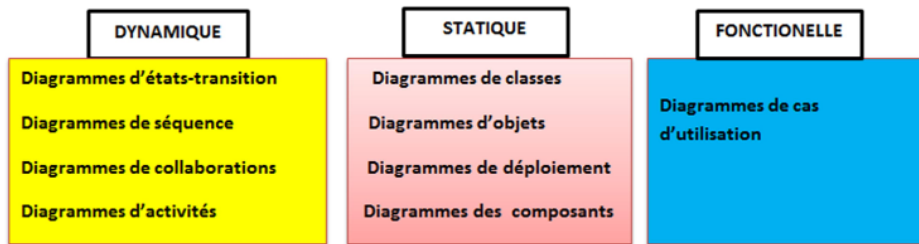


FIGURE 4.8 – Les diagrammes de modélisation par UML

4.4.2 L'architecteur de notre système :

L'architecteur de notre système est un architecteur trois tiers car il existe trois niveaux (trois couches) qui se communiquent entre eux comme suit :

- **L'utilisateur** : l'utilisateur qui formule et envoie une requête à destination du serveur d'application, il reçoit les résultats finaux de la recherche.
- **Le serveur d'application** : (appelé aussi middleware) le serveur chargé de fournir la ressource mais faisant appel à un autre serveur
- **Les serveurs secondaires** : fournissant un service au premier serveur, dans notre application ils sont les serveurs web (les sources qu'on a mentionnées précédemment).

4.4.3 Modélisation du système

Nous allons modéliser la conception de notre application en utilisant un diagramme de chaque type. Notre choix est pour :

- **l'aspect fonctionnel** : le diagramme de cas d'utilisation qui sert à montrer les grandes fonctionnalités de système.
- **L'aspect dynamique** : le diagramme de séquence pour définir les interactions entre les objets dans le cadre temporelle.
- **L'aspect statique** : le diagramme de classes pour représenter l'architecture interne du système.

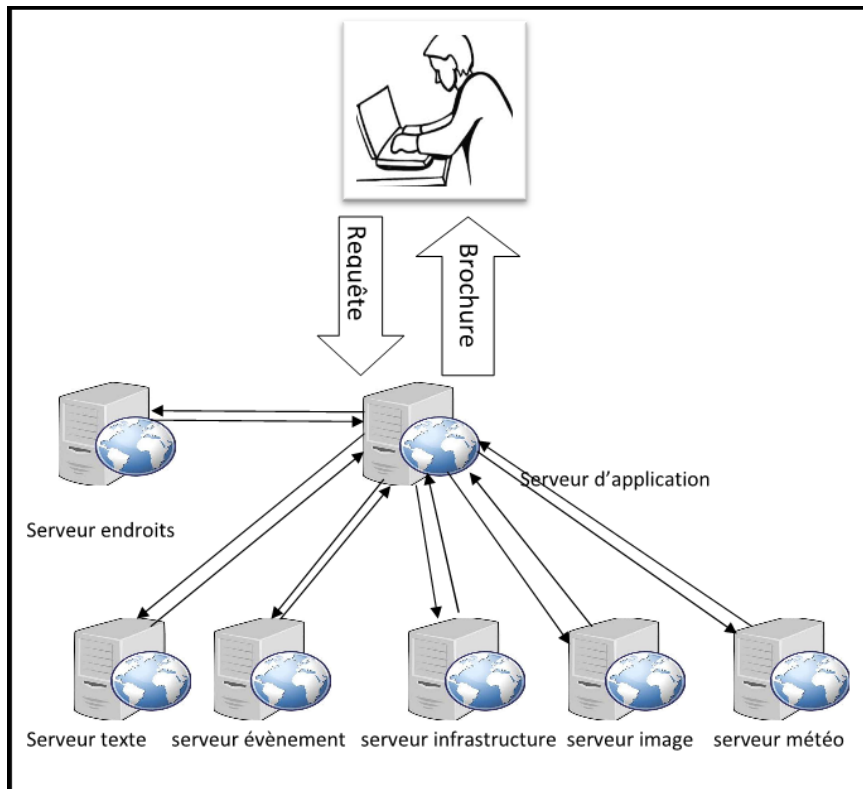


FIGURE 4.9 – l'architecture de notre système

Diagramme de cas d'utilisation :

Le diagramme de cas d'utilisation est utilisé pour donner une vision globale du comportement fonctionnel d'un système logiciel. Un cas d'utilisation permet de décrire les interactions d'un système avec son environnement. Dans un diagramme de cas d'utilisation, les utilisateurs sont appelés acteurs (actors), ils interagissent avec les cas d'utilisation (use cases).

- Identification des acteurs du système : Représente le rôle d'une entité externe qui interagit avec le système, ce rôle décrit les besoins. Dans notre cas les acteurs du système sont : **L'administrateur** : C'est le propriétaire du site, l'administrateur représente un rôle très important dans la gestion de notre système, cette gestion consiste en la mise à jour et la maintenance du site pour garantir un meilleur service aux internautes. **L'utilisateur** : ou bien l'Internaute c'est la personne qui effectue la recherche dans l'application pour avoir une idée et découvrir les différentes destinations à visiter.
- Identification des cas d'utilisation : Décrit ce qu'un système fait, ci-dessus le dia-

gramme de cas d'utilisation général de notre système :Décrit ce qu'un système fait, ci-dessus le diagramme de cas d'utilisation général de notre système :

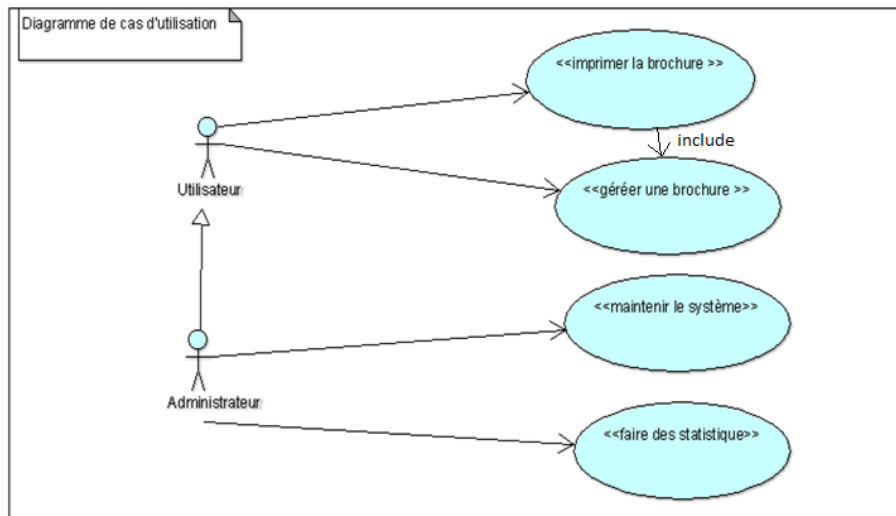


FIGURE 4.10 – Diagramme des cas d'utilisation général du système

Diagramme de séquence

Il représente séquentiellement le déroulement des traitements et des interactions entre les éléments du système et/ou de ses acteurs. Il permet également de montrer les interactions d'un système avec son environnement ainsi qu'il modélise un système de manière dynamique et il s'attache principalement à montrer la circulation et l'ordre chronologique des messages, autrement dit il décrit la circulation de l'information [29]. Ci-dessus les diagrammes de séquence qui modélisent notre système qui sont la génération de la brochure, impression de la brochure :

Diagramme de classe :

Ce diagramme utilise en génie logiciel la modélisation orientée objet qui s'appuie sur la représentation de modèle statique et dynamique. Nous allons dans ce qui suit, présenter d'une manière générale la structure statique de notre système en termes de classe et de relation. Dans la phase d'analyse, ce diagramme représente les entités (des informations) manipulées par les utilisateurs. Dans la phase de conception, il représente la structure objet d'un développement orienté objet

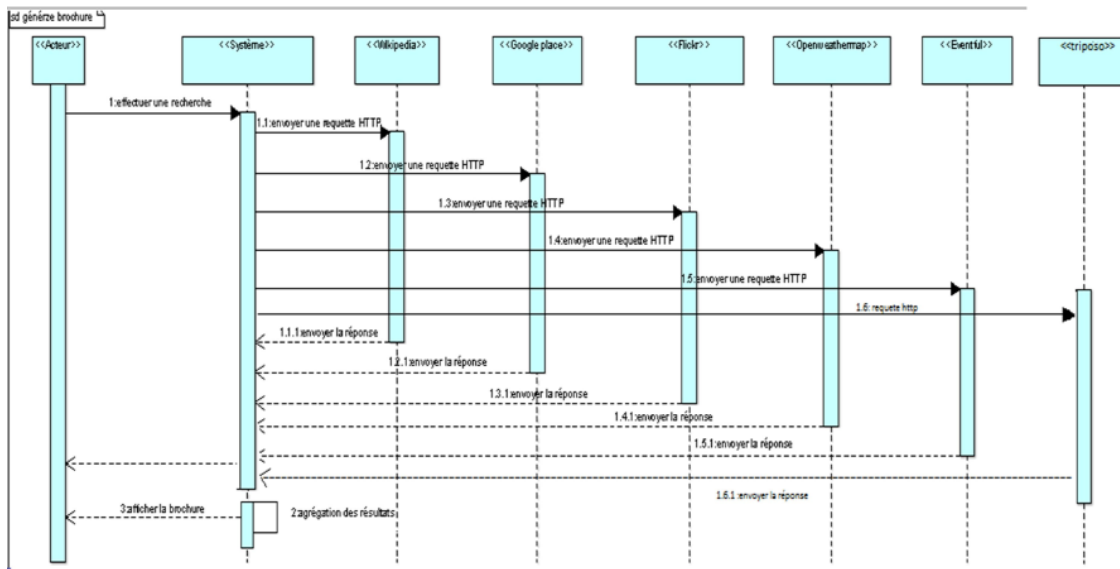


FIGURE 4.11 – Diagramme de séquence pour Imprimer la brochure

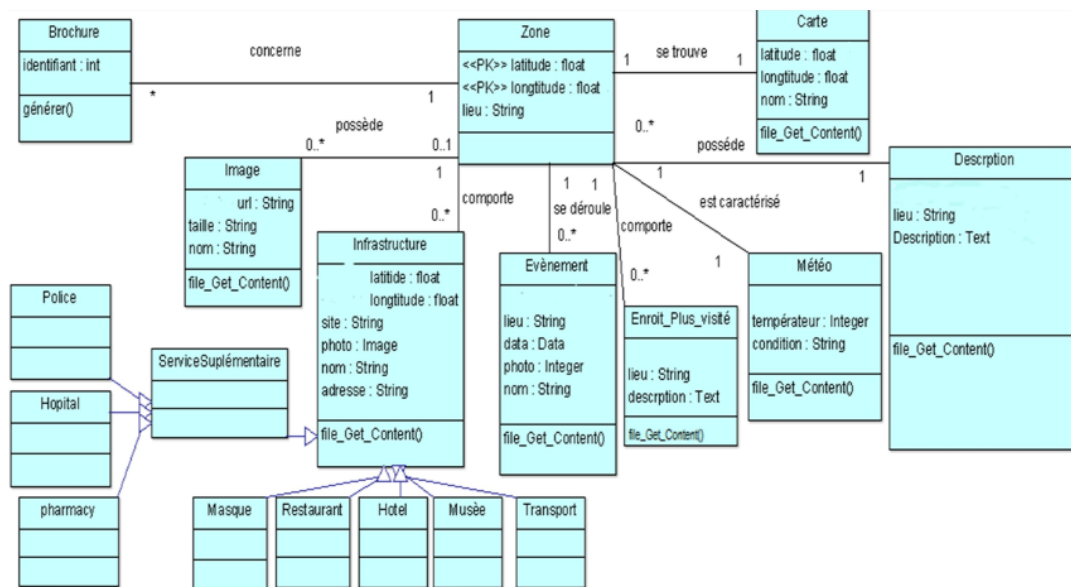


FIGURE 4.12 – Diagramme de classes

4.5 Conclusion :

Dans ce chapitre, on a proposé le contenu de notre brochure touristique, ainsi que les sources d'informations utilisées par notre système et la méthode d'agrégation pour construire la brochure. On a modalisé notre système par l'identification des diagrammes de cas d'utilisation, de séquence, et de classe. Le choix de la l'UML était car il nous servit à bien définir les besoins des clients, la cohérence des fonctionnalités et des données et la compréhension rapide du programme. Cela pour passer à la phase de l'implémentation et qui sera l'objet du chapitre suivant.

Réalisation et implémentation

5.1 Introduction

Ce chapitre consacré à détailler l'implémentation physique de notre système. L'objectif de cette phase est d'aboutir à une application finale, exploitable par les utilisateurs. Nous spécifierons dans ce qui suit les aspects techniques en présentant l'environnement de développement, les outils et les langages. Apres nous présenterons les interfaces offertes par l'application en utilisant les captures d'écran. En final nous conclurons notre chapitre.

5.2 Environnement de travail

5.2.1 Environnement matériel

Nous avons utilisé pour réalisé notre application une machine acer classique E1-531 i5 4GB, Intel Core i5-3210M.

5.2.2 Environnement logiciel et développement

Les outils et technniques

Les différents outils et techniques utilisés pour la réalisation de ce travail sont :

Les APIs : API est un acronyme pour Applications Programming Interface. Une API est une interface de programmation (code source) qui permet de se «brancher» sur une application pour échanger des données, autrement dit, L'objectif est de fournir un accès à

des services pour répondre à des requêtes qu'un autre programme informatique pourrait lui faire.

- **Openweathermap** : Open Weather Map fournit des cartes interactives des conditions météorologiques actuelles et historiques. L'API Open Weather Map gratuit permet aux utilisateurs de récupérer la météo actuelle dans une ville ou une station météorologique, les mesures historiques d'une station météorologique ou une liste de villes et / ou de stations météorologiques dans un rectangle donné (limité par des coordonnées géographiques). L'API utilise les appels RESTful émis au format JSON.
- **Google map** : L'API Google Maps permet d'intégrer Google Maps sur des pages Web de développeurs externes, à l'aide d'une interface JavaScript simple ou d'une interface Flash. L'API inclut la localisation de la langue pour plus de 50 langues, la localisation de régions et le géocodage, ainsi que des mécanismes pour les développeurs d'entreprise qui souhaitent utiliser l'API Google Maps sur un intranet. Les services HTTP de l'API sont accessibles via une connexion sécurisée (HTTPS) par les clients Google Maps API Premier
- **Google place** : l'api Google place permettre de renvoi plusieurs d'information sur un lieu selon les types existe comme des établissements ,...etc via une requête http en précise les coordonnes géographique la longitude et latitude de la place demandée .
- **Flickr** : L'API Flickr peut être utilisée pour récupérer des photos du service de partage de photos Flickr à l'aide d'une variété de flux - photos et vidéos publiques, favoris, amis, pools de groupes, discussions, etc. L'API peut également être utilisée pour télécharger des photos et des vidéos. L'API Flickr prend en charge de nombreux protocoles, notamment REST, SOAP, XML-RPC. Les réponses peuvent être formatées en XML, XML-RPC, JSON et PHP.
- **Eventful** : est la plus grande collection d'événements au monde, se déroulant sur les marchés locaux à travers le monde, des concerts et des sports aux événements individuels et aux rassemblements politiques.. L'API événementielle permet d'accéder à la base de données complète du site, permettant aux développeurs de l'intégrer et aux fonctionnalités de la plate-forme Eventful dans les applications Web. Les tâches API communes incluent la recherche d'événements et de lieux. L'API utilise

les appels RESTful et les réponses sont formatées en XML, JSON ou YAML

- **Triposo** : est une plateforme de contenu de voyage intelligente. Nous utilisons nos algorithmes intelligents pour parcourir le Web et analyser des millions de sites Web et de critiques. Notre application de guide de voyage est la meilleure vitrine de notre plateforme. Notre application vous permet de choisir vos hôtels, sites, activités et restaurants préférés et de les ajouter à votre liste de seaux. Vous pouvez ensuite réserver vos favoris en toute transparence via l'application. Toutes vos réservations et places enregistrées sont maintenant dans un endroit facile à trouver
- **Wikipedia** : Wikipedia est construit à l'aide de MediaWiki, qui à son tour prend en charge une API, et fait également. Cela fournit aux développeurs un accès au niveau du code à la référence entière de Wikipedia. Le but de cette API est de fournir un accès direct de haut niveau aux données contenues dans les bases de données MediaWiki. L'API prend en charge les clients JavaScript basés sur le Web, L'API utilise les appels RESTful et prend en charge une grande variété de formats, y compris XML, JSON, PHP, YAML et autres.

Les langages utilisés

- **HTML** : (HyperText Markup Language) : C'est le langage universel utilisé sur les pages Web lisibles par tous les Navigateurs Web (Internet Explorer, Netscape, Mozilla, etc....). Ce langage fonctionne suivant l'assemblage et la combinaison de balises permettant de structurer et donner l'apparence voulue aux données textes, images et multimédias suivant la mise en page voulue.
- **CSS** :Cascading Style Sheets (feuilles de styles en cascade) : servent à mettre en forme des documents web, type page HTML ou XML. Par l'intermédiaire de propriétés d'apparence (couleurs, bordures, polices, etc.) et de placement (largeur, hauteur, côte à côte, dessus dessous, etc.), le rendu d'une page web peut être intégralement modifié sans aucun code supplémentaire dans la page web. Les feuilles de styles ont d'ailleurs pour objectif principal de dissocier le contenu de la page de son apparence visuelle.
- **JavaScript** : (souvent abrégé JS) est un langage de programmation de scripts principalement utilisé dans les pages web interactives mais aussi côté serveur. C'est un langage orienté objet à prototype, c'est-à-dire que les bases du langage et ses

principales interfaces sont fournies par des objets qui ne sont pas des instances de classes, mais qui sont chacun équipés de constructeurs permettant de créer leurs propriétés, et notamment une propriété de prototypage qui permet d'en créer des objets héritiers personnalisés.

- **PHP** : HyperText Preprocessor, plus connu sous son sigle PHP, est un langage de programmation principalement utilisé pour produire des pages Web dynamiques via un serveur HTTP, mais pouvant également fonctionner comme n'importe quel langage interprété de façon locale. PHP est un langage impératif orienté-objet.
- **XML** : Extensible Markup Language (XML) est un format de texte simple et très flexible dérivé du langage SGML (ISO 8879). Initialement conçu pour relever les défis de l'édition électronique à grande échelle, XML joue également un rôle de plus en plus important dans l'échange d'une grande variété de données sur le Web et ailleurs.
- **Xpath** : Le langage XPATH offre un moyen d'identifier un ensemble de noeuds dans un document XML, XPath permet de parcourir un fichier XML d'une façon à la fois simple et puissante. De la sorte, en peu de temps, un développeur peut rapidement et aisément extraire les informations qui l'intéressent, même dans un document qui en comporte bien plus. Par exemple récupérer le contenu d'une balise précise ,ou récupérer du contenu en fonction de la valeur d'un attribut d'une balise, récupérer un ensemble de balises avec leur contenu et les parcourir
- **Bootstrap** : Bootstrap est une infrastructure de développement frontale, gratuite et open source pour la création de sites et d'applications Web. L'infrastructure Bootstrap repose sur HTML, CSS et JavaScript (JS) pour faciliter le développement de sites et d'applications réactives et tout-mobile. La conception réactive permet à une page ou une application Web de détecter la taille et l'orientation de l'écran du visiteur pour adapter automatiquement l'affichage ;

Environnement de développement :

WampServer : WampServer (anciennement WAMP5) est une plateforme de développement Web Open source de type WAMP, permettant de faire fonctionner localement (sans se connecter à un serveur externe) des scripts PHP. WampServer n'est pas en soi un logiciel, mais un environnement comprenant deux serveurs (Apache et MySQL), un interpréteur

de script (PHP), ainsi que phpMyAdm..etc.

5.3 Présentation de l'application

Cette partie sera essentiellement consacrée à la présentation des principales interfaces du système sous forme de capture d'écrans.

5.3.1 La page d'accueil :

La page d'accueil permet aux utilisateurs touristes d'effectuer une recherche d'un lieu par région, pour faciliter la recherche et éviter les erreurs de saisie on a utilisé l'auto-complète qu'on a implémenté pour qu'il affiche des suggestions de régions dès que l'utilisateur commence la saisie de sa requête.

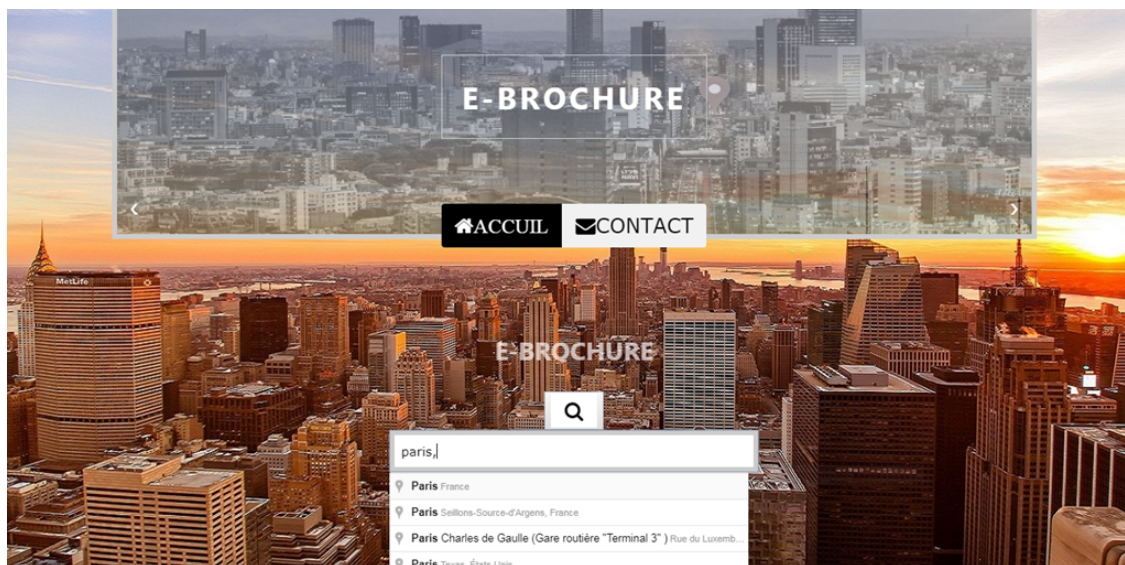


FIGURE 5.1 – Page d'accueil

5.3.2 La brochure

Les figures ci dessous illustrent les parties de la brochure générée



FIGURE 5.2 – Page récapitulant le contenu



FIGURE 5.3 – Description du lieu

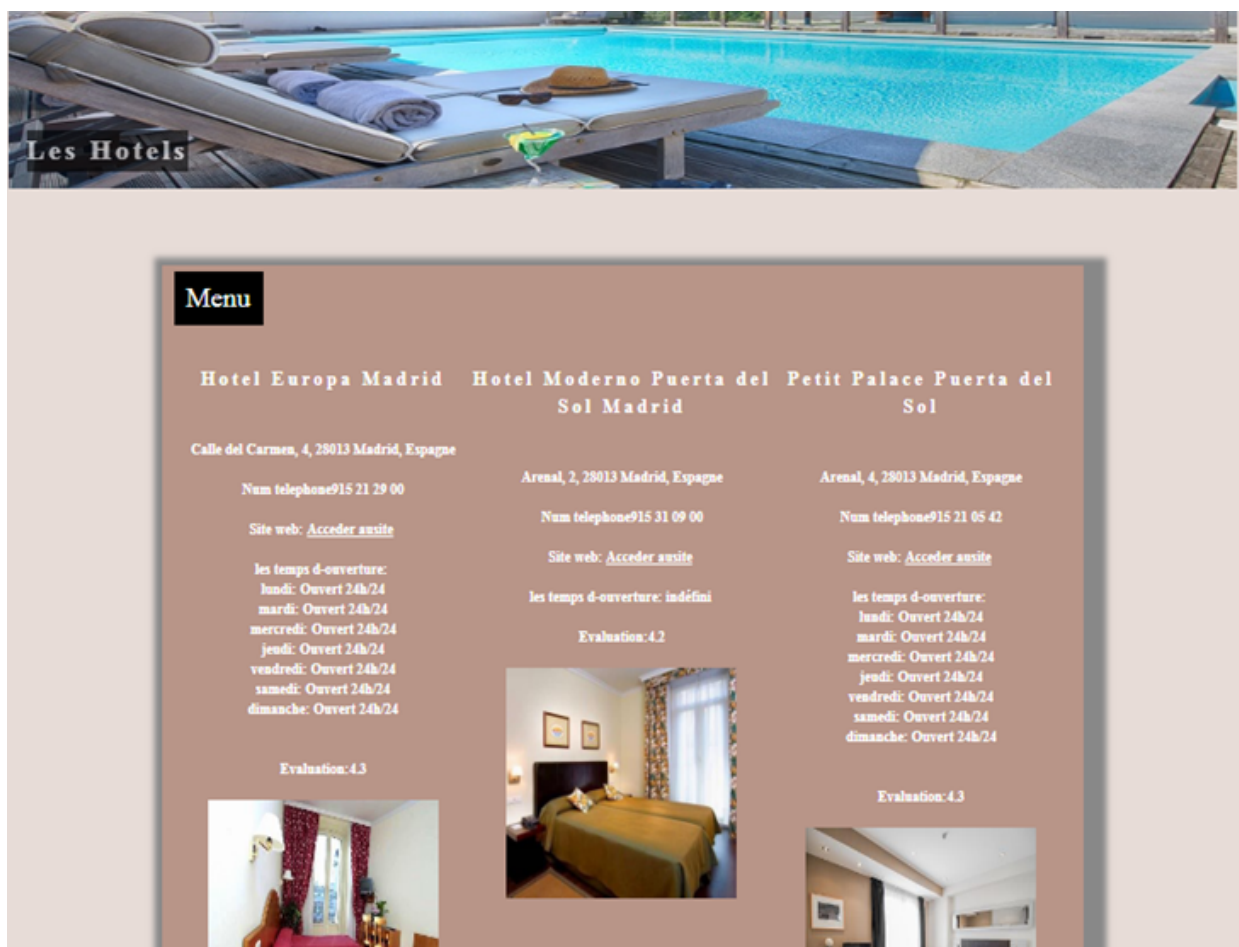


FIGURE 5.4 – Page des Hôtels

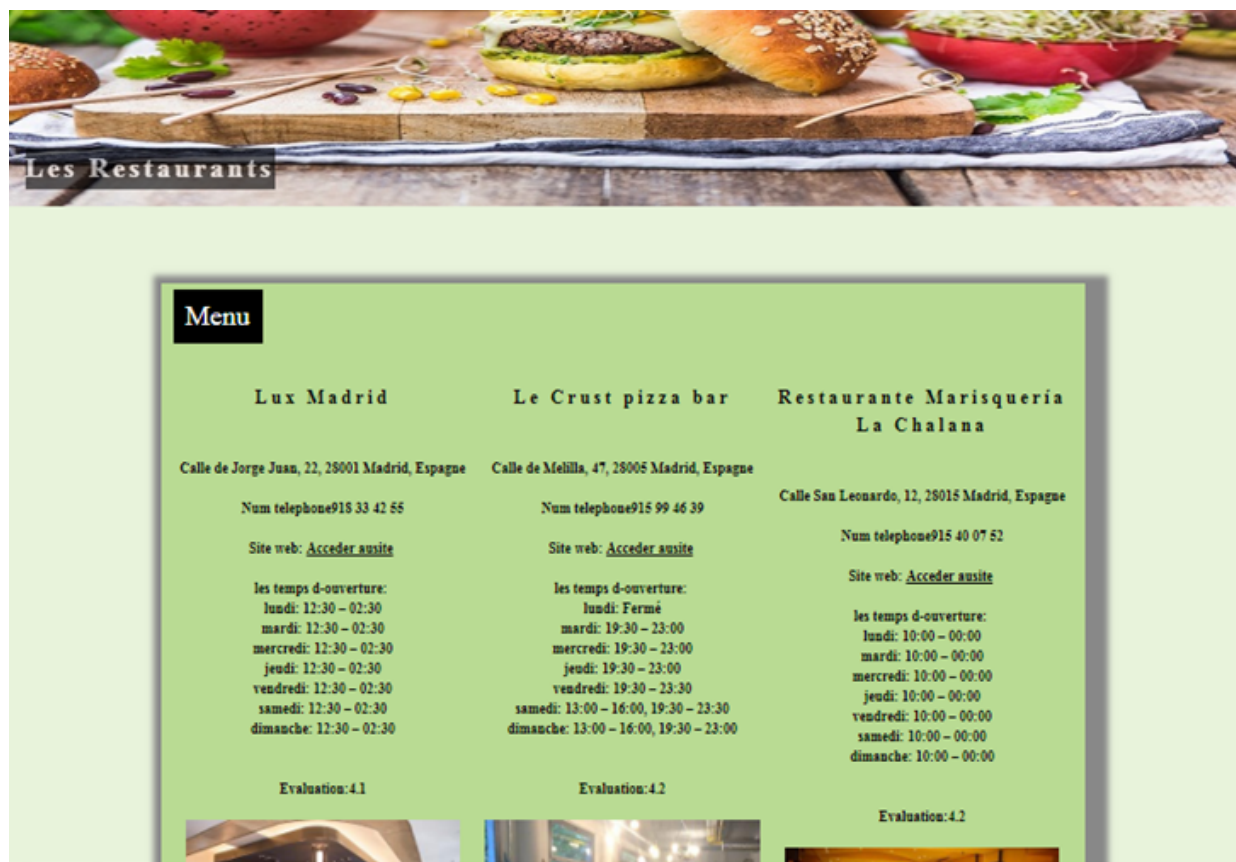


FIGURE 5.5 – Page des restaurants



FIGURE 5.6 – Page des événements

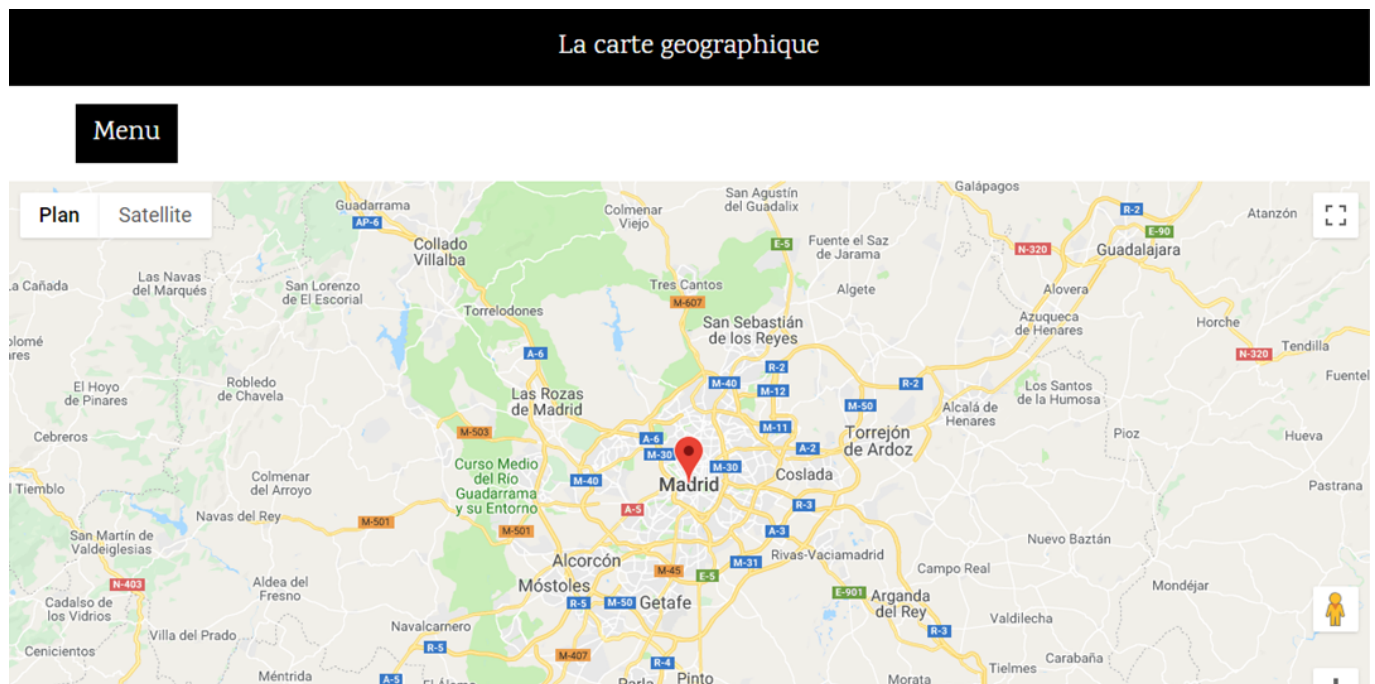


FIGURE 5.7 – Page de la carte géographique

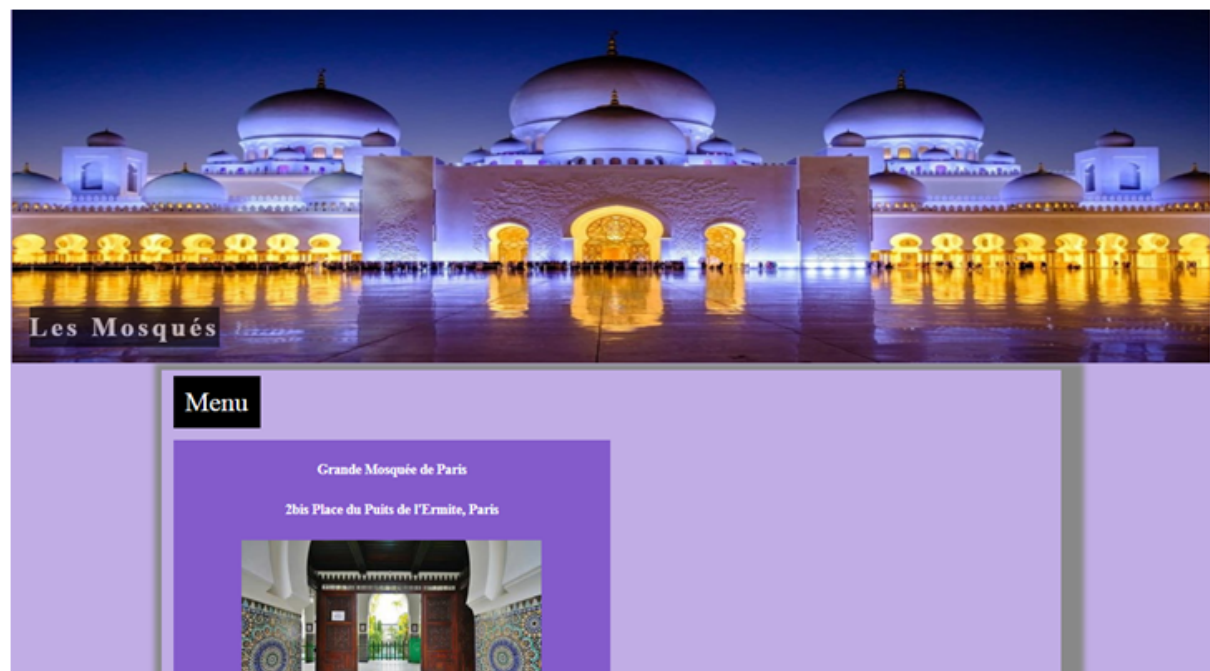


FIGURE 5.8 – Page des mosquées

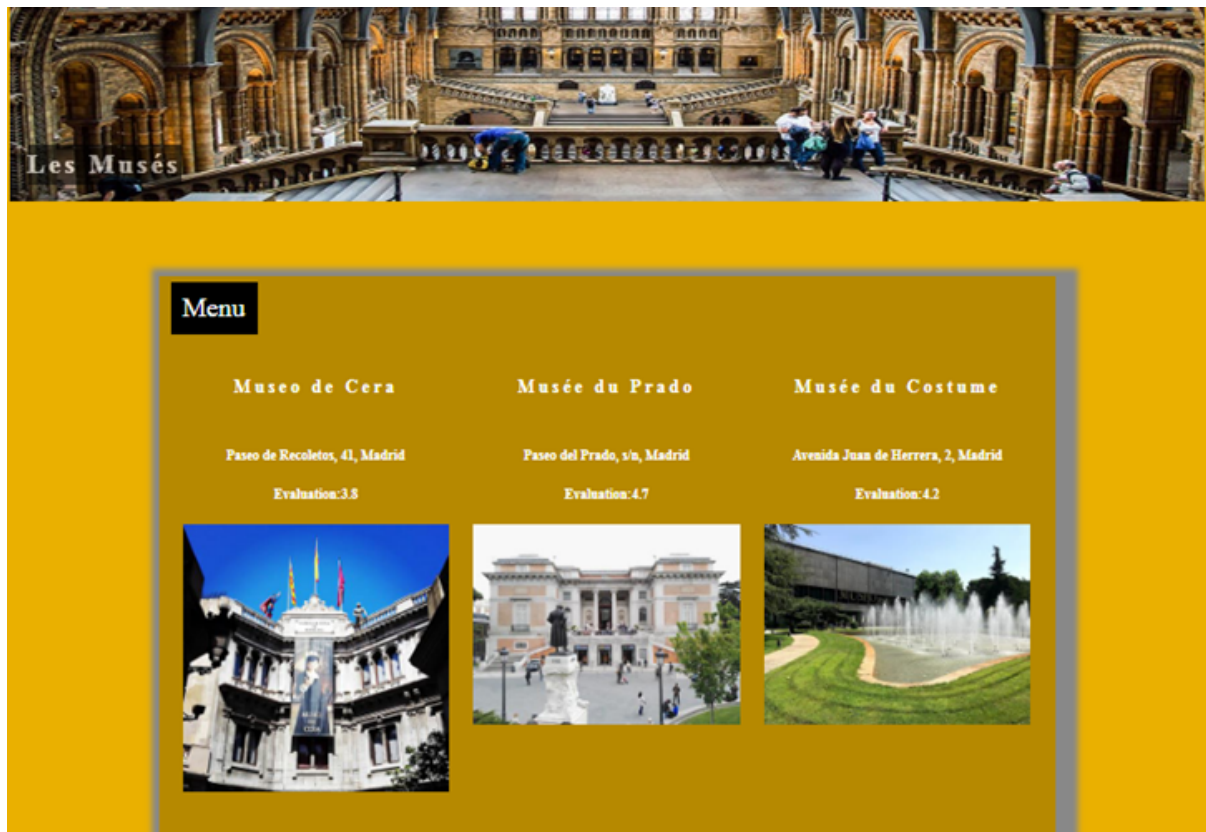


FIGURE 5.9 – Page des musés

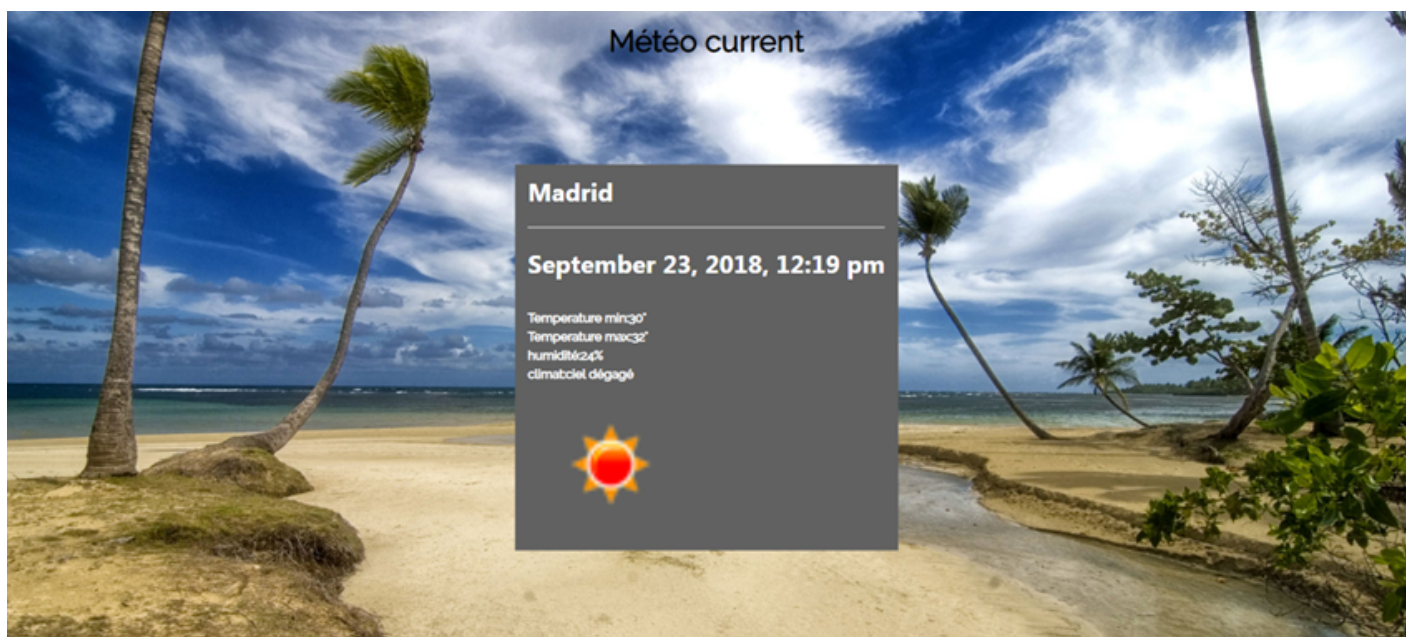


FIGURE 5.10 – Page des metéos

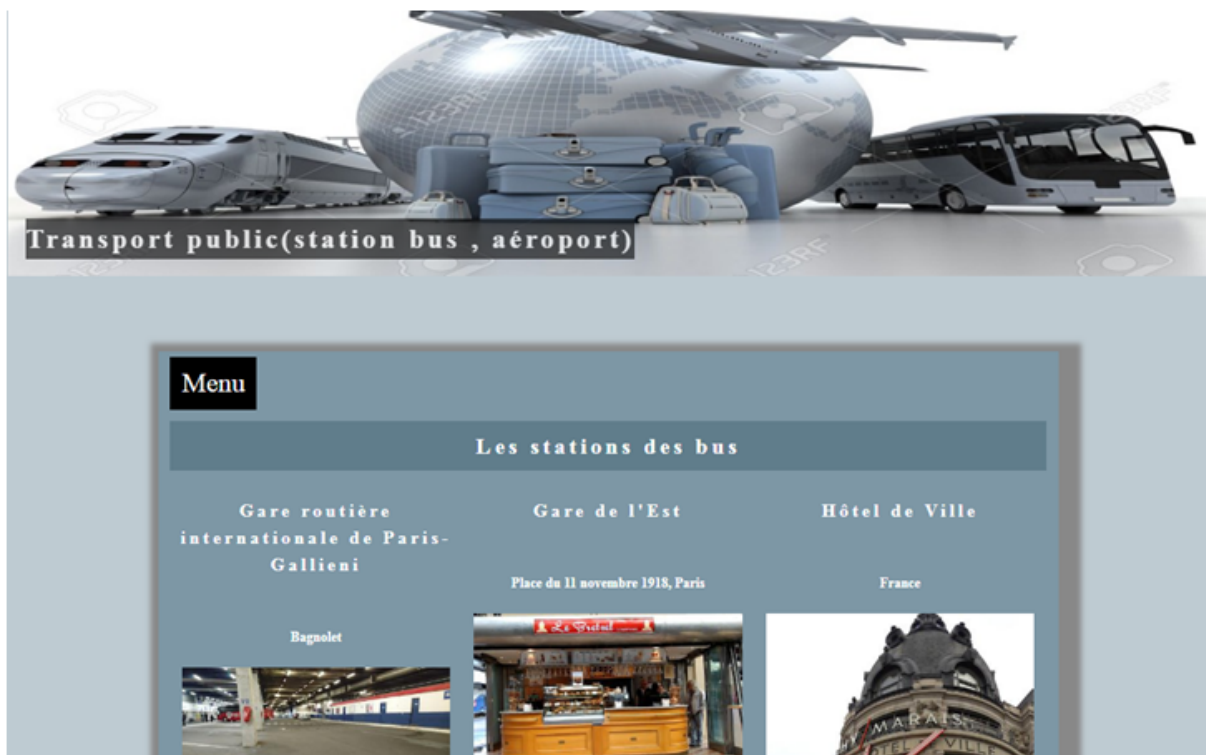


FIGURE 5.11 – Page des stations de transport

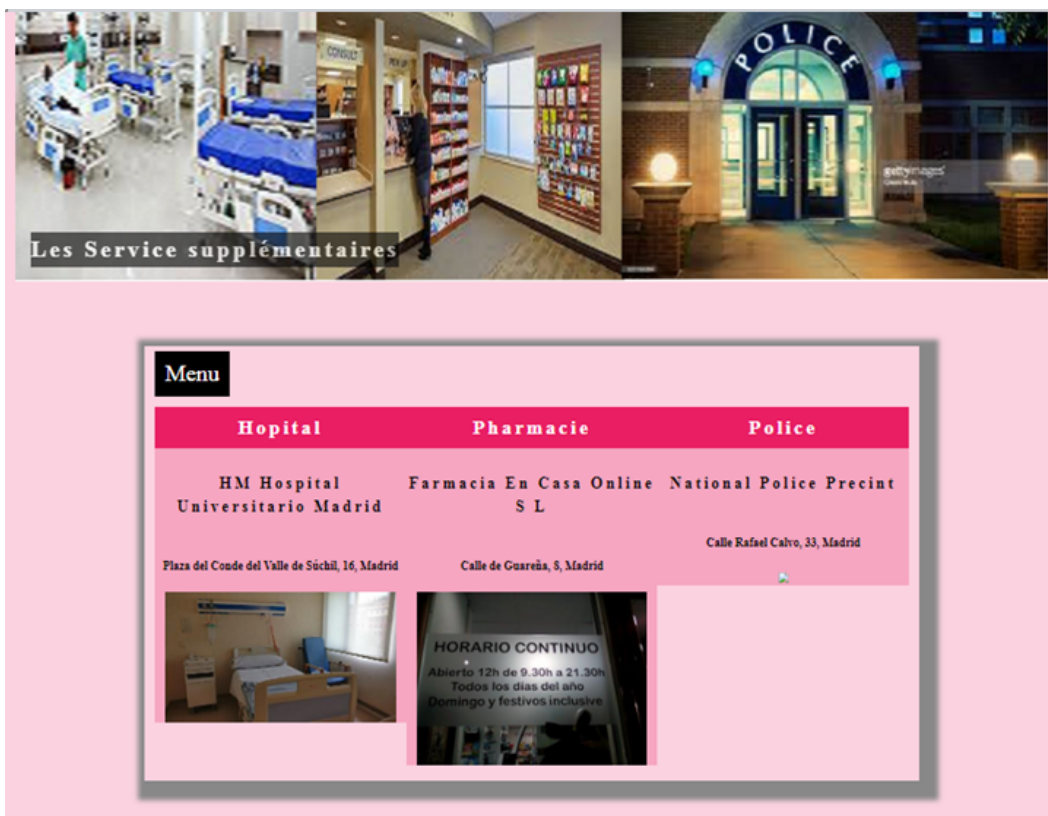


FIGURE 5.12 – Page des services supplémentaires

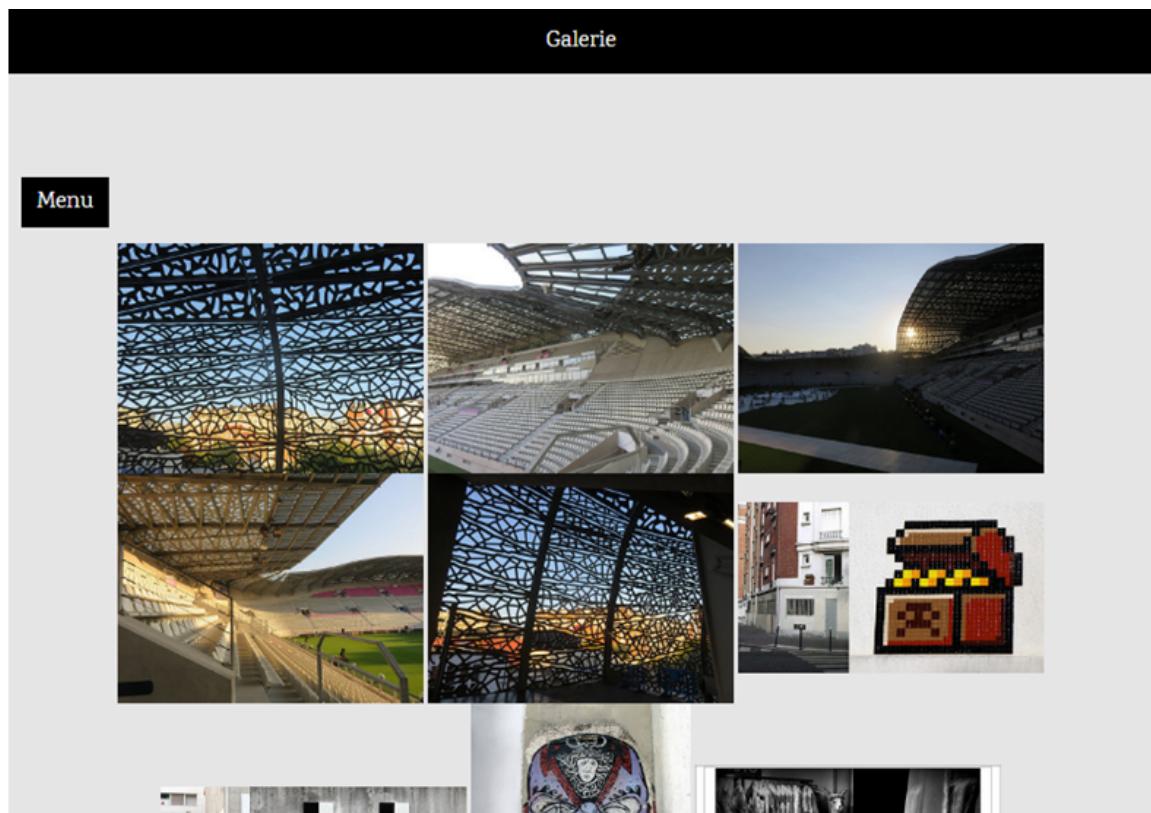


FIGURE 5.13 – Page de galerie d'images

5.4 Conclusion :

L'objectif de ce chapitre est l'implémentation des différentes fonctionnalités décrites dans le chapitre de conception, la représentation de ces fonctionnalités, et les outils et langages utilisés.

Conclusion générale

Les travaux décrits dans ce mémoire s'inscrivent dans le contexte général de la Recherche d'Information (RI) et plus particulièrement dans le cadre de la RI agrégée. La RI agrégée est un nouveau paradigme de recherche d'information qui, contrairement à la RI classique, ne se limite pas à renvoyer une liste de documents pertinents à l'utilisateur mais à produire automatiquement des espaces de réponses agrégé cohérent et bien organisés. Des unités d'information pertinentes (nuggets) recherchées dans plusieurs sources d'information sont sélectionnées et automatiquement agrégée pour produire des réponses complètes, bien organisées à l'utilisateur final. .

La recherche agrégée est un domaine de recherche récent, A. Koplaku [18] a proposé un cadre général et un schéma conceptuel pour le processus de recherche d'information agrégée qui est compatible avec toute différente approches liée à la RI agrégée. Ce processus est composé à trois phases : la répartition des requêtes (Query Dispatching (QD)), la recherche des unités d'information (Nuggets Retrieval (NR)) et l'agrégation des résultats (Result Agregation (RA)). La génération automatique de brochures touristiques constitue un très bon cas d'application du paradigme de la RI agrégée. En effet l'information touristique qui compose la brochure touristique est une information variées et diversifiée (Texte, photos, météo, infrastructure, ...). Ces information se trouvent sur des sources différentes et doivent être recherchées, sélectionnées et agrégées afin de produire une brochure touristique cohérente et bien organisée.

Nous avons dans ce modeste travail conçu et développer un système de génération automatique de brochures touristique en se basant sur les concepts et techniques liées au paradigme de la RI agrégée. Suite à une requête utilisateur qui se traduit par le nom d'une zone géographique ou d'une destination touristique, un ensemble de sources

d'information sont interrogées et l'information récupérée est organisée automatiquement afin de produire une brochure complète, riche et bien organisée sur la destination objet de la requête. .

Pendant l'implémentation de notre application nous avons rencontré un certain nombre de difficultés, notamment la celle liées aux API des sources d'information utilisées qui étaient payantes pour la plupart. Celles qui étaient gratuites n'offraient pas un nombre satisfaisant de résultats. Nous avons aussi été confrontés au manque de données concernant certaines régions. Nous avons tout de même essayé au maximum de satisfaire au mieux et répondre aux besoins des touristonautes.

Comme perspectives, à court terme nous souhaitons compléter notre application par d'autres fonctionnalités telle que le partage de la brochure et enrichir le contenu de la brochure avec d'autres sections. Nous envisageons aussi de programmer une version, android pour mieux satisfaire le demandeur de service.

Bibliographie

- [1] Hernandez N. *Ontologie de domaine pour la modélisation du contexte en recherche d'information*. PhD thesis, Université Paul Sabatier, 2006.
- [2] Salton. G, Fox. E, A, and H. Wu . Extended boolean information retrieval. *Communications of the ACM*, 26(12) :1022–1036, 1983.
- [3] Salton. G and M. McGill. Introduction to modern information retrieval. *McGraw-Hill Int. Book Co*, 1984.
- [4] M Daoud. *Accès personnalisé à l'information : approche basée sur l'utilisation d'un profil utilisateur sémantique dérivé d'une ontologie de domaines à travers l'historique des sessions de recherche*. PhD thesis, Université Paul Sabatier, 2009.
- [5] Vanessa Murdock and Mounia Lalmas. Workshop on aggregated search. *SIGIR Forum*, 42(2) :80–83, 2008.
- [6] ABDELKRIM BOURAMOUL. *RECHERCHE D'INFORMATION CONTEXTUELLE ET SEMANTIQUE SUR LE WEB*. PhD thesis, Université MENTOURI de Constantine, 2011.
- [7] SMAIL Nabila. *Contribution à l'analyse et à la recherche d'information en texte intégral : application de la transformée en ondelettes pour la recherche et l'analyse de textes*. PhD thesis, Université Paris-Est, 2009.
- [8] Tefko Saracevic. Relevance : A review of the literature and a framework for thinking on the notion in information science. part iii : Behavior and effects of relevance. *Journal of the American Society for Information Science and Technology*, 58(13), 2007.

-
- [9] M BOUCHER. Le rôle du contexte dans le jugement de pertinence en situation de repérage d'information. *Documentation et bibliothèques*, 2013.
- [10] Herzallah Abdelkarim. Recherche d'information. *cours MI-Umbb*, 2013.
- [11] HAJLAOUI Kafil. *Dispositifs de recherche et de traitement de l'information en vue d'une aide à la constitution de réseaux d'entreprises*. PhD thesis, Ecole Nationale Supérieure des Mines de Saint-Etienne, 2009.
- [12] Ronen GM, Penney S, and Andrews W. The epidemiology of clinical ,neonatal seizures in newfoundland. *a population-based study.J Pediatr*, 1999.
- [13] S. E Robertson. The probabilityrankingprinciple in ir. *Journal of Documentation*, 1977.
- [14] G SALTON. Smart retrieval system- experiments in automatic document processing. *systemeSMART(System for the Mechanical Analysis and Retrieval ofText)*., 1971.
- [15] Ba-Duy DINH. *Accès à l'information biomédicale : vers une approche d'indexation et de recherche d'information conceptuelle basée sur la fusion de ressources termino-ontologiques*. PhD thesis, Université Toulouse 3 Paul Sabatier, 2012.
- [16] N. Nassr. *Croisement de langues en recherche d'information : traduction et désambiguïsation de requêtes*. PhD thesis, Université Paul Sabatier, 2011.
- [17] A Kopluku. *Approaches to implement and evaluate aggregated search*. PhD thesis, Université Paul Sabatier, 2011.
- [18] Arlind K, Mohand Boughanem, and Karen S. Technical report-irit : Aggregated search : potential, issues and evaluation. 2009.
- [19] J. Callan. Distributed information retrieval. In *W. Croft (Ed.), Advances in Information Retrieval*, pages 127–150.
- [20] Bal K, Amghar Y, Réda Ghoamri, and H Mellah. Aggregated search techniques. *Usability.In Proceedings of the 2nd International Workshop on Web Intelligence I*, 2013.
- [21] Arugello J, R. Capra, and W. Wu. Factorsaffecting aggregated search coherence and search behavior. In *Proc. 19th Int. Conf.Information and Knowledge ManagementI*, 2013.
- [22] M. Lalmas. Advance topic in informationretrieval. *Springer Berlin HeidelbergI*, 2011.

- [23] Tim Berners-Lee. Worldwide web. *Proposal for a HyperText Project*, 2013.
- [24] Ali Tebbakh. *l'hypertexte au réseau sémantique, l'intérêt du wiki*. PhD thesis, l'Université de Lorraine, 2013.
- [25] J. C. Harlow Holloway. *Marketing for Tourism*. Essex : Pearson Education Ltd., 2004.
- [26] <https://www.holland.com/fr/tourisme/destinations/amsterdam.html>.
- [27] <https://www.visitbritain.com/fr/fr/angleterre/londres>.
- [28] Xavier Blanc and Isabelle Mounier. *UML 2 pour les développeurs : cours avec exercices corrigés*. Editions Eyrolles., 2006.
- [29] Olivier Glassey and Jean-Loup Chappelet. *Comparaison de trois techniques de modélisation de processus : ADONIS, OSSAD et UML*. PhD thesis, Institut de hautes études en administration publique, 2002.