République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Akli Mohand Oulhadj de Bouira

Faculté des Sciences et des Sciences Appliquées

Département d'Informatique

Mémoire de Master

en Informatique

Spécialité : *ISIL*

Thème

Systéme de Génération Automatique de revue de presse

Réalisé par

- Chellali Tarek
- Haddouche rabah

Devant le jury composé de :

- **Encadreur :**Kamel Bal
- **Président :**
- **Examinateur 1 :**
- **Examinateur 2 :**

2018/2019

Remerciements

Nous nous devons de remercier ALLAH le tout puissant pour toute la volonté et le courage qu'il nous a données pour l'achèvement de ce travail.

Aussi nous exprimons nos très sincères remerciements à nos encadreurs Mr Kamal Bal pour leur soutien, leurs conseils judicieux et leur grande bien vaillance durant l'élaboration de ce travail. Nous tenons à remercier les membres du jury, qui ont bien daigné siéger la soutenance de notre mémoire.

A nos chers professeurs du département informatique, un remerciement particulier et sincère pour tous les efforts que vous avez fournis pour nous encadrer tout au long de ces cinq années, vous nous avez enrichis avec vos connaissances et savoirs, nous avonsbeaucoupappris avec vous, vos remarques et conseils ont contribué à notre progression et amélioration au cours de notre cursus.

Nous voulons remercier très spécialement AISSA, qui a toujours été là pour nous.

Dédicaces

Je dédie ce mémoire à :

Ma mère, qui a œuvré pour ma réussite, de par son amour, son soutien, tous les sacrifices consentis et ses précieux conseils.

Mon père, qui peut être fier et trouver ici le résultat de longues années de sacrifices et de privations pour m'aider à avancer dans la vie.

Mes frères et sœurs qui n'ont cessé d'être pour moi des exemples de persévérance, de courage et de générosité.

À l'esprit du mon frère décédé le grand HADDOUCHE Rabah.

Toute ma famille, et mes amis, A mon binôme Tarek et à tous ceux qui ont contribué de près ou de loin pour que ce projet soit possible.

Je vous dis tous merci.

HADDOUCHE Rabah

Dédicaces

Avec un énorme plaisir, un cœur ouvert et une immense joie, que je dédie ce modeste travail à :

Ma Mère '' Tu m'as donné la vie, la tendresse, l'amour, et le courage pour réussir. En témoignage, je t'offre ce modeste travail pour te remercier pour tes sacrifices et pour l'affection dont tu m'as toujours entourée ".

Mon Père '' L'épaule solide, l'œil attentif compréhensif et le personne le plus digne de mon estime et de mon respect. Aucun dédicace ne saurait exprimer mes sentiment, que dieu te préserve et te procure santé et lange vie".

Ma chère famille : Hichem, Anouar Oussama, Rayan Saif El Islam, Loubna Hanan, pour leur encouragement contenu et leur soutient qu'ils trouvent l'expression de ma haute gratitude

Mon très cher amis et mon Binom Rabah et sa famille.
Mon très chers amis Ayoube et Abde el hak et Mohamed et leurs familles.

Mes chères amis et à ceux qui aimaient :Theldjoune Said, Khelfane Rabeh,Hocine Ahcen Boudissa_Ilyes, Brahim_Errebai, , Amayasse, Fateh, Malek, Djamel, Sami, Halim ,Kamel.

CHELLALI Tarek

Table des matières

Table des figures

Liste des tableaux

Liste des abréviations

URL	Uniform Resource Locator
RSS	Really Simple Syndication
XVIIe	dix septième
XVIIIe	Dix-huitième
PDF	Portable Document Format
AP	Associated Press
WWW	World Wide Web
RA	Recherche Agrégé
SIGIR	Special Intrest Group on Information Retrival
RI	Recherche d'Information
DUC	Document Understanding Conferences
TAC	Text Analysis Conference
NLG	Natural-language generation
TF	Term Frequency
IDF	Inverse Document Frequency
RAM	Random Accese Memory
HTML	Hyper Text Marckup language
CSS	Cascading Style Sheet
PHP	Hypertext Preprocessor
XML	eXtensible Marckup language
AJAX	Asynchronous JavaScript and XML
DOM	Document Object Model
JASON	JavaScript Object Notation
VSC	Visual Studio Code
API	Application Programming Interface

Introduction générale

Dés l'avènement d'Internet, les premiers journaux occidentaux (américains et européens) ont commencé la diffusion de l'information sur le web. Les premiers sites web des quotidiens d'information sont apparus en 1995. Rapidement, toutes les organisations de presse se sont tournées vers le média en ligne, Internet.

Dans les années 90, on a assisté à la prolifération de nombreux sites qui diffusaient des informations d'actualité sans avoir au préalable des liens avec le métier du journalisme. Dans la même décennie, plusieurs formes de sites d'informations sont nées, entre blogs, webzines (magazines en ligne) ou sites indépendants. De nos jours, et avec l'émergence des Technologies de l'Information et de la Communication et la démocratisation de la production de contenu sur internet, le nombre de sites d'informations et les quantités énormes de contenu créé ne cessent d'augmenter exponentiellement chaque jour. Des flux d'actualités sont continuellement diffusés via des centaines voir des milliers de sites spécialisés dans les news.

Avec toutes ces quantités informationnelles et ce nombre croissant de médias en ligne, le lecteur se trouve incapable de traiter ce flux d'informations manuellement, et même les outils technologiques (classiques) ne remédient pas à cette problématique. Le lecture n'arrive plus à suivre facilement l'actualité.

La revue de presse est une synthèse des sujets d'actualité traités dans les différents titres et organe de presse. Sa génération manuelle nécessite une connaissance en journalisme et un effort humain considérable. De plus avec le caractère continu des flux d'actualité, la revue de presse classique n'est pas automatiquement mise à jour.

Le travail présenté dans ce mémoire a pour objectif principal de concevoir et développer un système de génération automatique de revue de presse. Il s'agit donc d'interroger un ensembles de sources d'actualités, d'identifier les différents sujets traités et de présenter la revue générée sous une forme bien organisée et facile à exploiter aux lecteurs. Pour arriver à cet objectif nous exploiterons le paradigme récent en recherche d'information qui est la Recherche Agrégée. Nous considérons la génération automatique de revue de presse comme un cas d'application de ce paradigme.

Nous structurons la suite de ce présent mémoire comme suit :

— Le premier chapitre : **Revue de presse** : consistera en la présentation du domaine de la presse et des revues de presse. Nous parlerons aussi dans ce chapitre sur les agitateurs de news.

— Dans le deuxième chapitre, **Recherche d'information agrégée**, nous présenterons le paradigme de recherche d'information agrégée sur le quel nous nous somme basé pour le développement de notre système. Nous citerons donc la différence de ce paradigme par rapport à la recherche d'information classique et nous nous focaliserons sur les différentes techniques d'agréation utilisées pour la présentation des résultats de recherche.

— Dans le troisième chapitre, **système de génération de revues de presse**, il sera question de présenter les différents composants et étape de la conception de notre système. Chaque étape du processus de génération de la revue de presse est décrite et documentée et les choix de techniques ou d'outil est justifié.

— Dans le quatrième et dernier chapitre, **implémentation**, nous présenterons le produit final ainsi que sa réalisation et les différents outils et techniques qu'on a utilisé. Puis, avant de parler des perspectives possibles et de l'avenir de notre application, nous dresserons une conclusion générale du projet.

Chapitre 1

Presse et Revue de presse

1.1 Introduction

Le succès d'internet et de moteurs de recherche a bouleversé le monde des médias, et nombreux sont les journaux à proposer aujourd'hui une double édition papier/numérique, les revues de presse sont les méthodes qui permet de couvrir cette extension. Au cours de ce chapitre, nous parlerons du concept de revue de presse et de ses caractéristiques, types et règles et comment le réaliser, en suit on abordera les agrégateurs de nouvelles en général qui font un exemple de revues de presse.

1.2 Définition et Naissance de la presse

La presse est par définition est une un moyenne de communication et de transmission de l'information[1]. Elle fait donc partie des médias. d'une manière générale, la presse est l'ensemble des moyens de diffusion de l'information On distingue deux types de presse :

— la presse écrite (journaux, magazines...).

— la presse audio-visuelle (télévision, radio, internet...).

— la presse électronique (internet...).

La presse écrite est d'abord apparue sous différentes formes : les nouvelles qui étaient manuscrites, les occasionnels, les libelles, les placards, les almanachs. Souvent, il s'agissait de simples feuilles volantes. Cette presse plus ou moins clandestine était vendue en librairie et par colportage. Dès la Renaissance et aux XVIIe et XVIIIe siècles, une partie de l'information écrite se faisait par voie manuscrite, plus particulièrement dans le domaine

de la presse clandestine, mais non exclusivement. Ces ateliers de copistes, dont l'exemple parisien le plus célèbre reste la paroisse Doublet, produisaient des journaux que l'on nommait « nouvelles à la main ».[2]
Le premier périodique imprimé au monde, un hebdomadaire de quatre pages, titré Relation fut lancé à Strasbourg en décembre 1605 par Johann Carolus.[3]

1.3 Les différents types de presses

1.3.1 La presse écrite

La presse écrite désigne, d'une manière générale, l'ensemble des moyens de diffusion de l'information écrite, ce qui englobe notamment les journaux quotidiens, les publications périodiques et les organismes professionnels liés à la diffusion de l'information.

1.3.2 La presse audio-visuelle

La presse audiovisuelle est l'ensemble des médias qui permettent la diffusion de l'information soit par l'image ou par le son, soit uniquement par le son.

1.3.3 La presse électronique

La presse électronique est une forme de la presse utilisant Internet comme principal support, par le biais notamment de versions électroniques de médias traditionnels, ou bien de journaux en ligne.

1.4 Revues de presse

1.4.1 Définition d'une revue de presse

Une revue de presse est une production spécifique. Il s'agit d'un relevé de presse, d'une synthèse des titres de la presse généraliste ou spécialisée, de source principalement écrite (presse écrite ou en ligne).[4]

C'est une présentation conjointe et comparative de divers commentaires émanant de journalistes différents et concernant un même thème ou un même événement.

La revue de presse prend aujourd'hui différentes formes avec l'émergence du Web. Auparavant diffusée sous forme de dossier papier, de bulletin, ou exposée sur des panneaux, elle est aujourd'hui diffusée au format électronique et peut prendre la forme d'un blog, d'un PDF ou encore d'un flux RSS. Il existe également des revues de presse radiophoniques ou télévisuelles.

1.4.2 Types de revue de presse

Il existe deux types de revue de presse :

— Revue de presse papier.

— Revue de presse électronique.

On peut aussi parler de revue de presse généraliste ou revue de presse thématique (ou spécialisée) pour une classification selon l'étendu de l'actualité considéré dans la revue de presse.

1.4.3 Objectifs de revue de presse

— La revue de presse est un outil de veille pour une entreprise. Elle lui permet de se situer dans son secteur d'activité et d'utiliser à bon escient l'actualité pour prendre des décisions stratégiques.

— Une revue de presse montre comment les médias (journaux et magazines, radios, télévision...) traitent les faits et les présentent à leurs publics. A travers sa synthèse, elle leur permet ainsi de se tenir informés d'un sujet, d'un événement, d'un domaine, d'un secteur d'activité ou d'une organisation particulière, de suivre pas à pas toutes ses nouveautés et ses évolutions, de connaître ses échos dans la presse.

— L'intérêt d'une revue de presse est aussi de refléter le pluralisme des médias : variété des attitudes et des points de vue des médias face à l'actualité, diversité des opinions.

— L'objectif d'une revue de presse est de synthétiser les titres de presse en hiérarchisant l'information. L'objectif est également d'arriver à mettre en lumière la manière dont l'information est traitée en mettant par exemple le doigt sur des écarts de jugement, des contradictions entre différentes sources.

1.4.4 Caractéristiques d'une revue de presse

— La revue de presse peut être diffusée sur différents supports : presse écrite, radio, télévision, Internet...

— La revue de presse suit une logique particulière : ses informations sont classées , exposées selon un plan et un angle (problématique...).

— Les sources de presse citées dans la revue de presse peuvent être variées aussi bien par leur Périodicité (quotidienne, hebdomadaire, mensuelle, bimensuelle...), leur orientation politique que par leur zone de diffusion (départementale, régionale, nationale, internationale) et leur domaine de spécialisation (économie, politique, science...).... Et pour chaque information donnée, la revue de presse propose plusieurs sources et points de vue différents qui sont présentés soit en opposition, soit de façon complémentaire (précisions, nuances...).

1.4.5 Règles d'une revue de presse

Dans le domaine du journalisme, un certain nombre de règles doivent être respectée durant l'élaboration d'une revue de presse :

— Respect du droit de réciprocité : compilation de plusieurs sources sur un même sujet.

— Regroupement organisé des informations par rubriques, thèmes ou événements ; travail de création et de classement.

— Respect des droits d'auteurs (droit moral et patrimonial) : reprise d'articles ou d'informations, citations courtes qui ne devraient pas dispenser le lecteur de lire l'article original (Site Web : lien URL vers l'article original), mention complète de l'auteur et de la source permettant au lecteur de s'y reporter aisément.

— Non-dénaturation de l'esprit ou de la forme de l'œuvre citée. [4]

1.4.6 Comment réaliser une revue de presse ?

Nous allons désormais procéder pas à pas pour réaliser une revue de presse d'actualité papier et/ou électronique :

Le Sourcing

La première étape est de sélectionner les sources d'informations utiles.
Ces sources varient en fonction des besoins : il est possible de ne sélectionner que des quotidiens nationaux ou de mêler presse d'information, magazines et presse professionnelle en fonction du sujet traité et de la couverture géographique et chronologique.
Il est possible de sélectionner des rubriques ou des pages à l'intérieur même de ces sources, par exemple les pages culture et société des grands quotidiens d'information.

La sélection des articles

La presse papier : La sélection des articles papier, que l'on appelle également dépouillement en bibliothèque, s'effectue après avoir lu la presse en diagonale. Il faut porter une attention particulière aux titres, et aux intertitres.
Seuls les articles traitant spécifiquement du thème doivent être sélectionnés et le thème lui même.

La presse en ligne : La sélection d'articles en ligne s'effectue au moyen d'une veille médias. Cette veille peut s'effectuer en pull et/ou en push :

— Veille en pull : elle consiste à se rendre directement et régulièrement sur des sites d'information pour lire la presse en ligne et sélectionner les articles pertinents. Il est ainsi possible de consulter des sélections d'articles en ligne sur les sites des principaux journaux français.
— Veille en push : elle permet de s'abonner à des listes électroniques et des fils RSS ou de recevoir des messages d'alerte lorsque'un article intéressant est mis en ligne, grâce à des bases de données comme Factiva, des agrégateurs de presse comme Google Actualités ou encore des lettres d'information électroniques.

La constitution et la diffusion

La revue de presse papier : Il vous suffit de coller les photocopies des articles sélectionnés sur un support adapté en n'oubliant pas de mentionner les références bibliographiques complètes de chaque article. Il est recommandé de classer les

articles par thèmes ou rubriques ainsi que par ordre de parution (ordre chronologique).

La revue de presse en ligne : La revue de presse électronique ne consiste pas à copier et coller les articles sélectionnés, mais à présenter leurs références bibliographiques complètes (nom de l'auteur, titres de l'article et du journal, date de parution) en les accompagnants de leur adresse Web d'origine. Il est possible de ne citer que les premières lignes de l'article pour donner une idée de son sujet tout en respectant le droit d'auteur.

La revue de presse électronique doit également être classée par thèmes ou rubriques et respecter l'ordre de parution des articles.

Plusieurs moyens s'offrent à vous pour diffuser :

— La diffusion par mail : en n'oubliant pas de convertir votre fichier au format PDF pour qu'il ne soit pas modifiable par d'autres.

— La diffusion sur un blog : il vous est également possible de déposer votre revue de presse sur un blog personnel ou étudiant.

1.4.7 Example des revues de presse

Ci-dessus des captures d'écrans d'exemples de revue de presses

FIGURE 1.1 – Captures d'écrans des exemples de revues de presses

1.5 Les agrégateurs de nouvelles (news aggregators)

1.5.1 Définition des agrégateurs de nouvelles

Un agrégateur est une entité qui regroupe plusieurs grandeurs ou flux en un seul. Dans le domaine informatique, un logiciel agrégateur tresse plusieurs fils de syndication en même temps. Il prévient de la mise à jour de sites web ou des actualités qu'ils publient et importe le contenu nouveau en question.

Les agrégateurs de nouvelles peuvent porter des noms différents. Ils sont parfois appelés lecteurs de nouvelles, lecteurs de flux, agrégateurs de contenu et lecteurs RSS, mais quel que soit leur nom, ils fonctionnent tous selon le même principe de collecte des informations et de leur classement dans un endroit facile d'accès[5]. Les sites d'agrégation de nouvelles prennent des histoires sur tout le Web et les placent sur une seule page Web, généralement regroupées par sujet ou par catégorie. Bien que cela puisse sembler au premier abord comme si l'agrégateur volait du contenu sur un autre site Web, il peut s'agir d'une relation mutuellement bénéfique. Le site d'agrégation de nouvelles reçoit du trafic des visiteurs qui le fréquentent, tandis que les sites Web d'origine des reportages sont plus exposés et donc plus fréquentés. Ceci est particulièrement utile lorsque le site d'agrégation est populaire et envoie les visiteurs vers des sites qui ne reçoivent normalement pas autant de trafic. Tant que le site d'agrégation de nouvelles fournit un lien et une attribution appropriée, aucun problème ne devrait se poser.

1.5.2 Les avantages d'agrégateurs de nouvelles

Le flux d'informations ne s'arrête jamais. Chaque jour, de plus en plus de données sont générées sur Internet, et certaines de ces informations affectent votre organisation. Les dirigeants d'entreprise doivent se tenir au courant des informations pertinentes dans leurs secteurs respectifs, mais ce n'est pas toujours facile à faire. Les agrégateurs de nouvelles vous permettent de lire les dernières nouvelles qui ont

un impact sur vous et votre entreprise, le tout à partir d'un emplacement unique. Il n'est pas nécessaire de parcourir des dizaines de sites Web dans l'espoir de trouver les informations que vous souhaitez. Les agrégateurs de nouvelles s'occupent du travail chargé et mettent cette information à portée de main.

D'un point de vue utilisateurs, les principaux avantages des agrégateurs sont les suivants :

— Economiser de l'argent : de nombreux médias essaient de limiter l'accès à leur contenu en utilisant un abonnement payant. Cependant, cela implique un certain nombre d'effets négatifs. Il est donc plus facile de devenir un agrégateur qui a déjà réglé tous les problèmes avec les titulaires du droit d'auteur.

— Précision accrue des résultats de recherché.

— Filtrage et catégorisation du contenu.

— Un point d'entrée.

— Nombre limité d'annonces : en échange, nous pouvons obtenir un texte pur et clair et les informations requises.

— Confort : il n'est pas nécessaire de parcourir d'énormes quantités de texte ; il suffit d'obtenir l'essence de tout le matériau et de décider quelles choses méritent notre attention .

1.5.3 Le fonctionnement d'un agrégateur de nouvelles

Aux débuts de l'agrégation de nouvelles, des conservateurs particuliers effectuaient des recherches sur le Web pour placer des articles sur leurs propres sites Web d'agrégation de nouvelles. Certains fonctionnent toujours manuellement, mais la grande majorité des logiciels et des applications d'agrégation de nouvelles fonctionnent via des systèmes et des processus codés. La plupart suivent un flux de travail similaire pour gérer toutes les nouvelles histoires Web ajoutées à Internet chaque jour.

La première partie de ce flux de travail traite de la collecte de données à partir de sites Web préalablement déterminés. Cela implique l'exploration d'une page Web, parfois en vérifiant les flux RSS pour collecter ces données. La partie suivante extrait des articles de ces données collectées. C'est la partie du flux

de travail où la pertinence des articles est déterminée. Vient ensuite l'étape de regroupement des articles en fonction des sujets abordés. De nombreuses méthodes différentes peuvent être utilisées à cet effet. Parmi les approches les plus courantes figurent les techniques basées sur les mots clés et la modélisation de sujets.[6]

Après cela, un article est choisi qui représente le mieux un groupe d'articles basé sur un sujet. C'est généralement l'article que l'agrégateur affiche en premier. Enfin, les sujets des articles sont ensuite visualisés afin que l'utilisateur puisse y accéder facilement. La plupart du temps, l'utilisateur peut déterminer les sujets qui l'intéressent le plus afin que l'agrégateur de nouvelles puisse se concentrer sur ces sujets.[7]

1.5.4 Les types de ressources d'agrégateur de nouvelles

Services de presse : Les médias (imprimés, radiodiffusés et en ligne) tirent une grande partie de leurs nouvelles de ces services, tels que Reuters ou Associated Press (AP), ce qui évite aux dépositaires individuels d'envoyer leurs propres journalistes partout. Les services sont si largement utilisés que vous devrez peut-être consulter plusieurs médias pour avoir une vision différente d'un événement ou d'une situation.

Sites de journaux : De nombreux journaux imprimés ont également leurs propres sites Web. Ils varient en fonction de la quantité d'informations qu'ils fournissent gratuitement.

Bases de données de nouvelles : Recherchez le contenu actuel, récent et historique des journaux dans des bases de données fournies gratuitement par les bibliothèques.

Sites d'actualités : Bien que les informations diffusées (à la radio et à la télévision) soient généralement consommées en temps réel, ces organisations proposent également des archives d'actualités sur leurs sites Web. Cependant, tous leurs articles ne sont pas fournis par leurs propres journalistes : certains proviennent des services de presse.

Blogs : Parfois, ce sont de bonnes sources pour les dernières nouvelles, ainsi que des commentaires sur l'actualité et les bourses d'études. Les auteurs qui écrivent plus objectivement ailleurs peuvent partager plus de points de vue et d'opinions, davantage de questions initiales et de conclusions sur une étude avant d'être prêts à publier des données définitives et des conclusions sur leurs recherches.

Journalisme citoyen : Un nombre croissant de sites s'adressent aux membres du grand public qui souhaitent rapporter des nouvelles de dernière heure et soumettre leurs propres photos et vidéos sur un large éventail de sujets. Les personnes qui font cela sont souvent appelées journalistes citoyens.

Fil d'actualité : Vous pouvez recevoir régulièrement des mises à jour sur des sujets spécifiques ou une liste des principaux titres afin que vous ne soyez pas obligé de visiter des sites ou de rechercher de nouveaux contenus sur un sujet.

1.5.5 Quelques agrégateurs de nouvelles

Il existe beaucoup d'agrégateurs de nouvelles disponible sur le net, donc on va mentionner quelque exemple comme :

— Google Actualité (Google News)
— Bing News
— NewsEola ou NewsMap
— Morning Briefing du New York Times
— Yahoo Actualité
— Techmem
— Orange Actualité

Nous allons dans la suite nous focaliser sur Google news et le presenter en detail

1.6 Google Actualité (Google News)

1.6.1 Présentation

Google Actualités ou Google News est un service en ligne gratuit de Google qui présente de façon automatisée des articles d'information en provenance de sources

sur le Web. Il fonctionne de la même manière qu'un moteur de recherche, en n'indexant que les articles de presse. Ce service, disponible dans vingt-deux pays, a été créé en avril 2002. La partie française du service est sortie de son statut bêta le 14 mai 2009 et le 25 janvier 2006 pour la partie américaine. En juin 2017, Google change le design de Google Actualités pour une version plus épurée et plus claire pour l'utilisateur. Elle propose également des articles dont les faits sont vérifiés dans la version américaine.[8]

FIGURE 1.2 – L'interface de Google Actualité

1.6.2 Fonctionnement général

Collecte de l'information

Afin de proposer une information qui ne fait appel à aucun journaliste, Google Actualités sélectionne des sources d'information dans les pays (ou les langues) où le service est proposé. Google Actualités recense par exemple 500 sources francophones et Google News 4 500 sources anglophones. Chacune des sources sélectionnées est scrutée plusieurs fois par jour par un robot Google : un Google bot, qui indexe les nouveaux articles. Le contenu des articles est ensuite analysé à l'aide d'algorithmes pour déterminer le sujet.

Ensuite les articles sont triés par degré de pertinence. Pour cela, Google prend en compte à la fois la notoriété de la source et compare avec d'autres articles publiés sur le même sujet, dans un même laps de temps. En fonction de cela, l'article va être plus ou moins bien classé sur la page de Google Actualités et lors des requêtes faites par les utilisateurs.

En 2009, Google News a commencé à répertorier également les pages de Wikipédia. Il s'agit cependant d'une phase de test uniquement aux États-Unis. Cette phase de test n'a cependant pas abouti à une implantation définitive.[9]

Diffusion

Sur la page principale, les articles sont regroupés parmi les rubriques suivantes : « À la une », « International », « France », « Entreprises », « Science/Tech », « Sports », « Culture et Santé ». Une mise à jour de la page est effectuée toutes les quinze minutes (option paramétrable). Il est possible de créer une page personnalisée avec ses propres rubriques : « Actualités personnalisées » comprenant les actualités relatives à des mots-clés, mais aussi des localités, le tout, à l'aide d'un compte Google.[10]

Grâce à la technologie de recherche de Google, il est possible de remonter plusieurs années en arrière pour rechercher dans les archives de Google Actualités, et de calibrer sa recherche sur des périodes très précises.

Evolution de Google Actualité

Le 28 octobre 2011 a lieu la mise en place d'une nouvelle interface pour Google Actualités avec un ajout notable de la catégorie « Les choix des rédactions » présentant les cinq articles proposés par une sélection de média français : L'Express, Le Monde, Le Figaro, Libération et Ouest-France.[11]

Le 28 juin 2017 a lieu la mise en place d'une nouvelle interface pour Google Actualités. Cette dernière semble soulever un mécontentement des utilisateurs. La polémique serait essentiellement due à son manque de compacité (notamment sur la version Ordinateur de bureau ou Desktop) ainsi qu'à la disparition d'options de personnalisation et de recherche.

Fin 2018, Google présente une version audio de Google Actualités destinée à son assistant personnel (Google Assistant). L'objectif est d'offrir un service de radio personnalisé à l'utilisateur. En décembre 2018, ce service n'est disponible qu'en version anglaise.[12]

1.7 Conclusion

Dans ce chapitre nous avons passé en revue sur le domaine de revue de presse, nous avons, particulièrement, introduit des notions de base tells que, la naissance de la presse, les types et règles de revue de presse. Nous avons aussi donné quelque exemple de revues de presse. Ensuit, nous avons parlé au mode général sur les agrégateurs de nouvelles. Dans le prochain chapitre, nous donnerons un bref aperçu sur le paradigme de recherche d'information agrégé.

Chapitre 2

Recherche d'Information Agrégée

2.1 Introduction

Nous allons tout au long de ce chapitre présenté un nouveau paradigme de recherche d'information qui a été développé ces dernières années. La présentation de ce paradigme est justifiée par le fait que nous nous baseront sur cet outil dans le développement de notre système de génération de revue de presse.

Nous savons que le succès des moteurs de recherche été initialement grâce à la recherche web, mais récemment de nombreuses recherches verticales deviennent populaire. Ils se concentrent dans le processus de la recherche sur un sujet précis (l'information juridique, l'information médicale, le sport, la finances . . . .), ou un type de médias (vidéo, images, nouvelles, blogs, . . .), ou une situation géographique, l'industrie et ainsi de suite.

Contrairement à la recherche d'information classique (horizontale) sur le web qui vise à l'exploration et l'indexation de la totalité du WWW, les recherches verticales tentent d'analyser et indexer seulement les pages web liées à leurs domaines prédéfinis. La plupart de ces recherches verticales (cartes, blogs, images, vidéos, . . .) peut maintenant être facilement accessible à partir des moteurs de recherche.

Cependant, pour améliorer la recherche, il y a une tendance aujourd'hui d'intégrer directement les résultats de recherche verticale au sein des résultats de la recherche web. Cela s'appelle la recherche agrégée (RA). Cela permet aux utilisateurs d'inter-

roger plusieurs sources (moteurs de recherche verticaux) à partir d'une seule interface. L'une des questions de la RA est de prédire quelles sources sont susceptibles de fournir des éléments pertinents aux requêtes. Par exemple la requête « Photos de la ville d'Alger » peut être résolue en effectuant une recherche d'images, d'autres requêtes peuvent être traitées à partir de plusieurs sources, « Visiter Le tassili » peut être satisfaite par des pages web ainsi qu'avec des images, des vidéos ou des cartes.

Le domaine de la recherche d'information agrégée a pris trois directions principales. Une direction étudie les interfaces adéquates pour la visualisation des résultats de la recherche agrégée, une autre direction implique des techniques qui peuvent prédire si les résultats d'une source doivent être inclus ou non en réponse à une requête donnée, ceci peut également être visualisé comme une forme de sélection de la source d'extraction de l'information distribuée.

La dernière orientation concerne l'évaluation des intérêts (utilité) de la RA ainsi que la définition des méthodologies pour évaluer les approches de la RA.

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

Lors de la conférence « **Special Interest Group on Information Retrieval** » SIGIR 08 [13], un atelier était tenu en particulier pour la recherche agrégée et dans lequel elle a été défini pour la première fois : *c'est une tâche cherchant à ressembler des informations provenant de sources différentes, et à les présenter dans une seule interface.* En d'autres termes, la recherche agrégée tente d'identifier le contenu nécessaire, de l'organiser et de le présenter à l'utilisateur de manière à faciliter sa recherche d'information. Un moteur de recherche agrégé est construit en surcouche d'un ou plusieurs autres moteurs de recherche (sources). Des informations de différents types (images, vidéo, etc.) et de différentes granularités (passage de texte, entités, attributs, etc.) sont reliées et parfois même combinées par une ou plusieurs relations logiques (association, groupe, ordre, etc.) afin de composer un résultat agrégé [14].

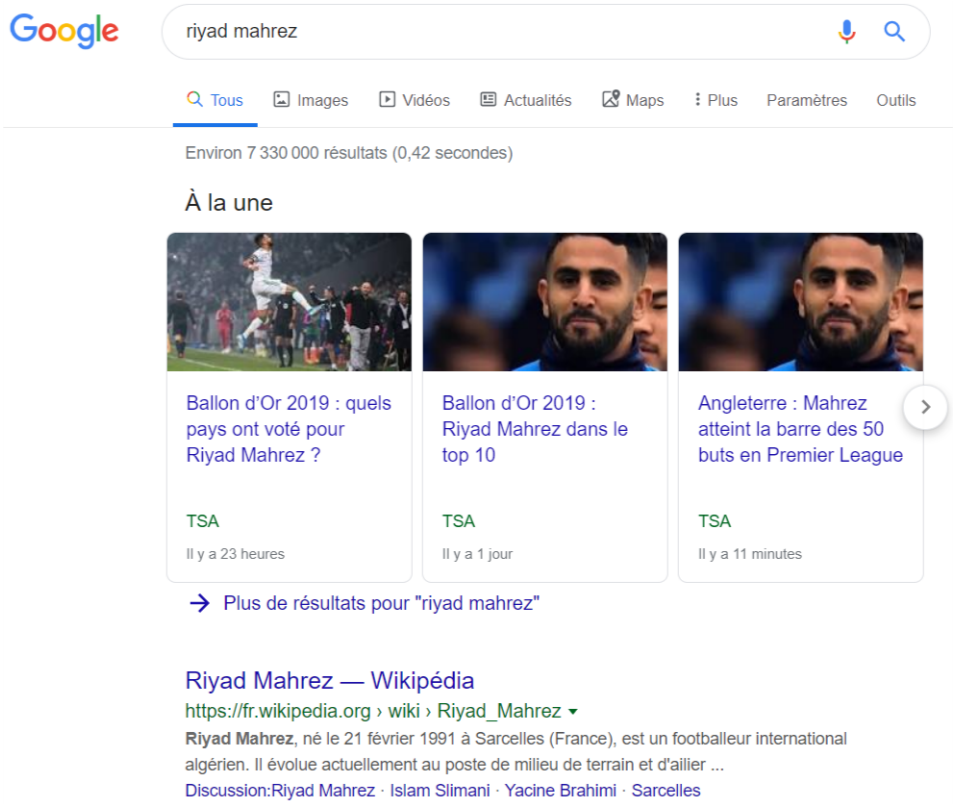


FIGURE 2.1 – Exemple de recherche agrégée – Universel Search

2.3 Processus générique de la RI agrégée

Nous savons que le processus de recherche d'information classique, appelé processus en U qui permet de mettre en relation une requête d'une part et une collection de documents d'autre part. Ce processus s'arrête à la récupération des résultats à partir de la collection et les présenter sous forme d'une liste de résultats à l'utilisateur. Le processus de recherche d'information agrégée ne doit pas s'arrêter là mais doit aussi prévoir une étape d'agrégation et de présentation des résultats issus de sources différentes. Arlind Kopliku (Kopliku, 2011) a proposé un processus générique pour la RI agrégée composé des phases suivantes :

— **Dispatching de la requête :** Cette étape, précédant la recherche elle-même, consiste à interpréter la requête et à sélectionner la ou les sources à interroger. Le terme source ici réfère à un moteur de recherche d'information indexant au moins une collection et utilisant un algorithme de recherche donné. La notion de source est importante,

car les systèmes de recherche d'information peuvent être classifiés en systèmes monosource et multisource. Quand plusieurs sources sont présentes, cette étape permet donc de sélectionner les sources à utiliser.

— **Recherche de nuggets :** La recherche agrégée permet de retrouver du contenu de différents granularités (phrase, passage, document) et de différents types (texte, image, vidéo,. . .). Ces contenus sont appelés nuggets, ou un nugget est un granule d'information de taille quelconque satisfaisant un besoin d'information. Cela peut- être un document entier ou une partie du document, mais aussi un simple mot. La recherche de nuggets correspond à l'identification de contenus pertinents, mais n'implique pas leur assemblage.

— **Agrégation des résultats :** Étant donné un ensemble de nuggets pertinents, l'agré- gation de résultats consiste à les assembler pour former la réponse finale. En fonction des systèmes de recherche agrégée, il peut s'agir de résumer, grouper, fusionner, trier, clusterier. . .

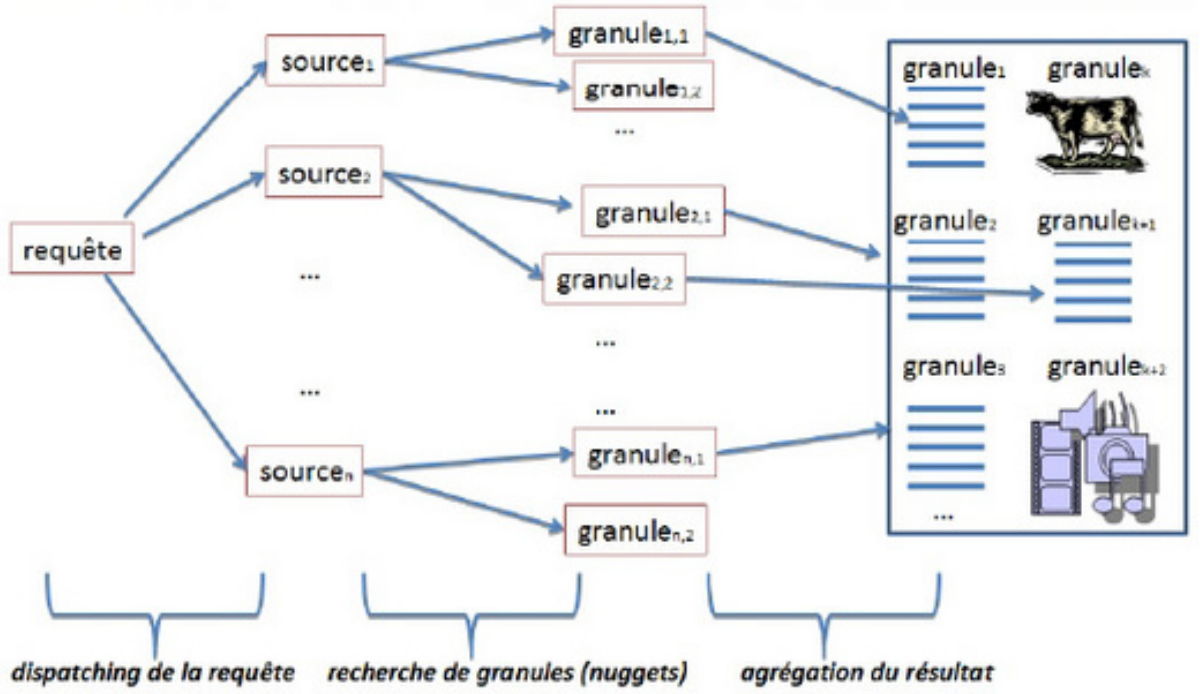


FIGURE 2.2 – Processus de recherche d'information agrégée

Les solutions commerciales comportent déjà des fonctionnalités de la recherche agrégée que nous pouvons trouver dans des contextes spécifiques tels que la recherche de produit ou recherche d'endroits. Par exemple, Wize.com offre une recherche de produits en donnant des résultats agrégés de plusieurs sources. La recherche locale de Google ajoute des numéros de téléphone, des images et pages web lorsque disponibles en plus de la carte résultat. L'agrégation semble être la tendance dans la recherche sur Web, cette tendance souvent désigné comme recherche unifiée, mais d'autres noms sont utilisés tels que la recherche verticale mélangée ou la recherche universelle. La nouvelle approche ajoute à la liste des pages Web présentées dans la page principale des résultats de la recherche un contenu vertical tel que des cartes, des images, des nouvelles,etc.

2.4 Structure d'agrégat

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. La recherche agrégée doit traiter l'organisation du contenu pertinent. Il n'est pas suffisant de collecter (trouver) 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 visualisation, et les préférences dans la visualisation du document agrégé [15].

Les exemples suivants illustrent des structures partielles de certaines informations agrégées :

— 3 images, 2 vidéos

— Une liste de 4 news, 3 informations supplémentaires.

— Paragraphe A, paragraphe B, paragraphe C, avec l'ordre de visualisation : A, B, C.

2.5 Les problématiques liées à la recherche d'information agrégée

Comme on l'a mentionné au niveau de l'introduction, ce nouveau paradigme de la recherche d'information a soulevé un certain nombre de problématiques supplémentaire par rapport aux problématiques classique de la RI agrégée :

— **Au niveau de la Granularité des résultats :**(ou l'identification du type de la réponse) le contenu des réponses renvoyées aux requêtes peut être différent. Pour certaines requêtes, une seule unité d'information suffit comme réponse, d'autres demandent de multiples unités.

— **identifier les unités d'information les plus pertinentes :**en RI agrégée, nous pouvons récupérer des unités d'information avec des granularités différentes et de types différents. Cela permet d'avoir une réponse finale plus exhaustive. Il n'est pas triviale d'identifier les unités qui devraient être utilisées pour composer la réponse finale. Quand devrions-nous utiliser une unité d'information au lieu d"un document entier ? Quand devrions-nous utiliser le contenu multimédia (images, vidéos, etc.) ? Quand devrions-nous utiliser les moteurs de recherche spécialisés (recherche d'images, recherche de vidéos, etc.) ? Ces questions sont les plus difficiles dans ce domaine.

— **Assembler les différents unités d'information dans un document cohérent :**la RI agrégée peut impliquer toutes les manières possibles d'assembler les résultats de recherche. Cela peut être un résumé, deux images et une définition, une table relationnelle, etc. L'un des objectifs de la RI agrégée est de choisir la meilleur agrégation selon les résultats de recherche disponibles. 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.

— **Redondance :** Bien que cette caractéristique est évidente dans les grands collections de données, un document agrégé parfait ne doit pas comporter des données redondantes afin d'exploiter l'espace de visualisation au maximum ainsi que mieux satisfaire l'utilisateur

— **Ambiguïté :** la multitude de type d'entité (ex : Washington peut être in-

terprétée comme une ville ou une personne) et de catégorie de documents (ex : politique, finance, . . . etc.) peuvent créer une ambiguïté sur le résultat le plus pertinent par rapport à la requête, car on peut trouver deux documents différents en réponse à la même requête.

2.6 Les approches d'agrégation

Les systèmes de recherche d'information fournissent aux utilisateurs une grande quantité d'information, d'où le défi de présenter efficacement les informations pertinentes à l'utilisateur. Lors de l'utilisation d'un moteur de recherche d'information pour effectuer des recherches dans des ressources électroniques, des requêtes simples retournent souvent de nombreux documents, dont beaucoup non pertinents à la recherche prévue. Par exemple, il existe plusieurs million de documents sur le World Wide Web concernant Michael Jordan, mais la plupart de ces documents portent sur la star du basket. Il est donc difficile de trouver des informations sur la personnalité de la télévision, le musicien de Jazz, le mathématicien, ou les nombreux d'autres personnes qui partagent ce nom.
De plus, il est bien connu que dans le contexte de la recherche web, les utilisateurs accèdent généralement à un très petit nombre de documents dans l'espace de résultats, fréquemment sur les trois premières pages, d'où l'importance de retourner des résultats plus diverses sur ces pages afin de fournir une bonne couverture de l'information disponible sur le web concernant les sujets de la requête. Plusieurs approches d'agrégation des résultats de recherche ont été développées.

2.6.1 Approche d'agrégation par Clustering

Cette approche consiste à regrouper les documents après récupération sous forme de clusters, et présenter un résumé de chacun de sorte que l'utilisateur peut choisir son groupe d'intérêt. Cette approche été proposée par Zeng et al qui considèrent que le regroupement des résultats de recherche dans des clusters permet d'avoir des documents qui se concentrent sur certains aspects de la requête. Parmi les systèmes basés sur cette technique, nous trouvons Clusty, QCS (Query, Cluster, Summarize) qui effectue les tâches suivantes en réponse à une requête :

— Récupère les documents pertinents.

— Sépare les documents récupérés en groupes par sujet.

— Crée un résumé pour chaque cluster.

D'autres exemples de systèmes de recherche d'information employant des algorithmes de Clustering pour organiser les ensembles des documents récupérés comprennent : Velocity/Clusty (Vivisimo, 2006), Infonetware / RealTerm (Infogistics, 2001), WiseNut (LookSmart,2006), Accumo (Accumo, 2006), iBoogie (CyberTavern, 2006), et le KartOO et systèmes UJIKO (KartOO, 2006). Ces systèmes organisent les documents en clusters et gé- nèrent une liste de mots-clés associés à chaque cluster. Les deux derniers systèmes pré- sentent également des représentations graphiques des clusters résultants. Comme avec le système de recherche ci-dessus, ces systèmes présentent aussi des extractions de document contenant un ou plusieurs termes de la requête, mais le seul résumé présenté est la liste de mots-clés.

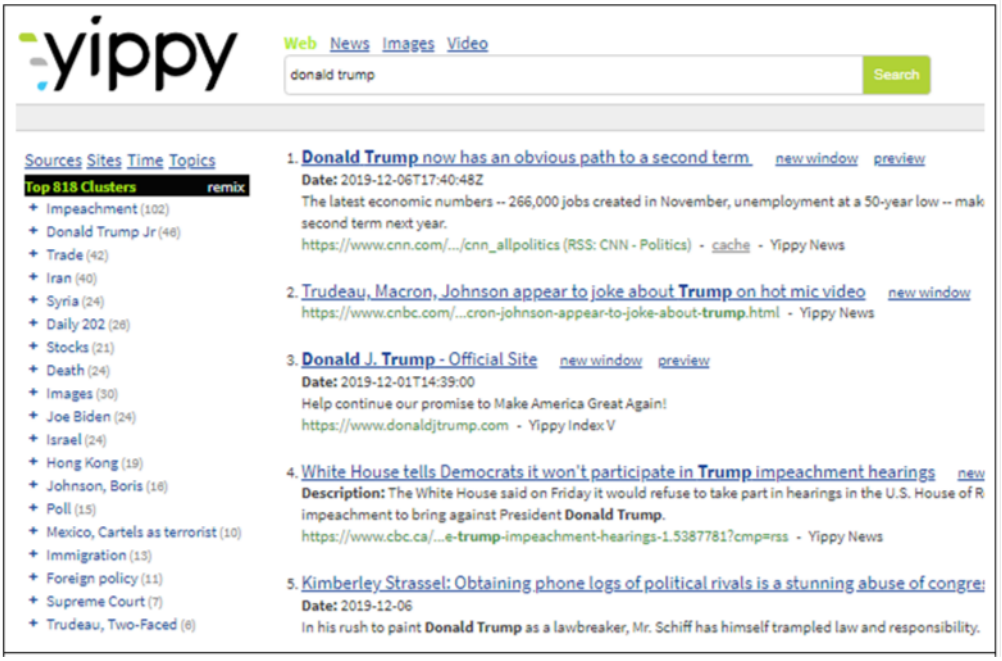


FIGURE 2.3 – Exemple de résultats de recherche du moteur Clusty/yippy

2.6.2 Agrégation par résumé multi-documents

La tâche générique de résumé multi-documents consiste à produire un résumé unique à partir d'une collection de documents qui relèvent probablement de la même thématique. Les premiers systèmes de résumé automatique multi-documents ont été développés par Kathleen R. McKeown et Dragomir R. Radev dans les années 1990. La tâche de résumé multidocuments s'est beaucoup inspirée des tâches correspondantes des conférences DUC et TAC. Ainsi, des avancées énormes ont été réalisées avec l'élimination de la redondance, principalement avec les travaux de Carbonell et Goldstein en 1998 [16].

Le résumé multi-document a été utilisé dans WebInEssence. En tant que technique d'agré- gation. WebInEssence est un système personnalisé de filtrage d'information et de résumé multi documents. Il est conçu pour aider les utilisateurs à trouver des informations utiles dans les documents sélectionnés sur la base de leurs profiles. Les résultats sont présentés sous forme de résumés de documents produits automatiquement. Cette technique est intéressante pour atténuer le problème de la surcharge cognitive d'informations et aider les utilisateurs à trouver l'information dont ils ont besoin. Le résumé automatique de document est le processus de sélection des informations les plus significatives d'un document.Lorsque l'entrée se compose de plus d'un document, on parle de résumé multi-documents.

FIGURE 2.4 – Interface du système Multi-source News

2.6.3 Agrégation relationnelle

D'autres travaux ont exploité les éléments de structures contenues dans les pages Web (telles que les tableaux et les listes) pour construire des résultats de recherche agrégés. Dans [17], les auteurs ont mis au point une technique d'agrégation basée sur la recherche d'attributs pertinents. Un résultat agrégé tabulaire de la forme "attribut / valeur" est construit pour chaque requête à travers trois étapes :

— la sélection des entités et des attributs pour la classe d'entité ou l'entité désignée par la requête,

— le filtrage des attributs récupérés

— le tri et le choix des attributs pertinents.

Dans la même catégorie de techniques, Google Labs a lancé un outil expérimental, appelé Google Squared, qui génère une réponse sous forme d'un tableau descriptif pour une requête donnée. La figure 2.5 montre un exemple de résultats Google Squared pour la requête «Technaute».

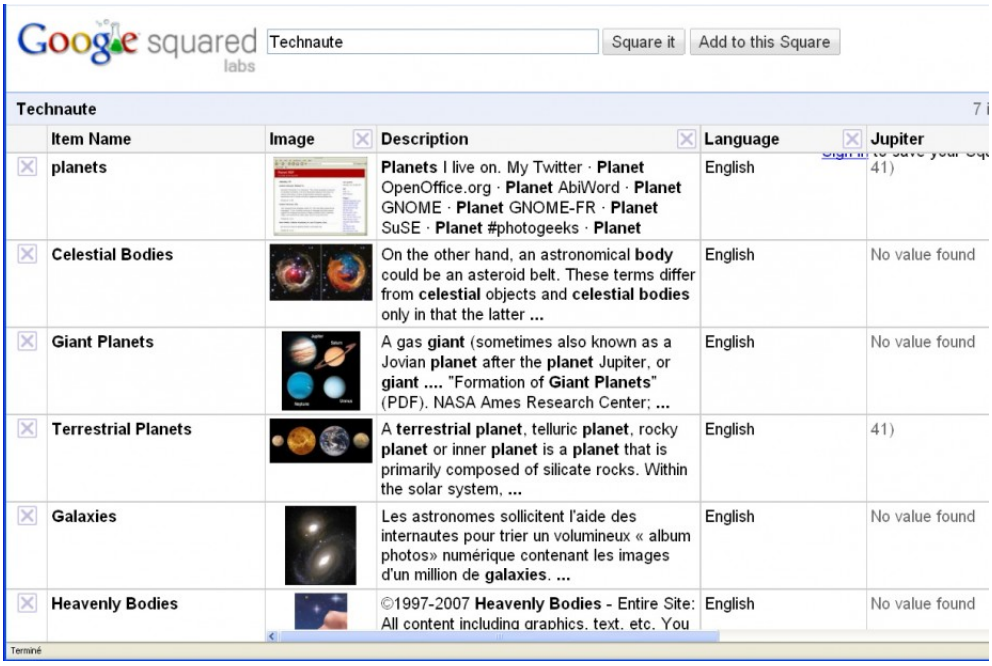


FIGURE 2.5 – Exemple de résultats Google Squared

2.6.4 Les vues agrégées « Aggregated views »

Les techniques d'agrégation précédemment mentionnées essaient de construire ou de générer une réponse (un document) à partir de différentes sources. La technique d'agrégation par vues agrégées quand à elle fusionne les différents résultats des sources dans une page de résultats unique. Cette technique est utilisée en recherche agrégée sur le web. Il est donc question de présenter la page de résultats de recherche comme une vue agrégée des différents résultats de recherche. Dans ce cas, chacune des sources d'information renvoie un seul type de média (vidéo, image, texte ...) ou de contenu (news, blog, météo, produits ...). Il existe deux types de vues agrégées : "blended views" et "non blended views". Dans les "blended views", les résultats de recherche hétérogènes provenant des différentes sources sont fusionnés et présentés verticalement dans une liste unique. Google universal utilise ce type d'agrégation. D'autres moteurs de recherche comme Yahoo Alpha 2 et Ask3D 3 utilisent les "non blended views" où chaque type de résultats est affiché dans une partie séparée (panneau) de la page de résultats de recherche.

Blended views

L'intégration Blended qui est appliquée par Google universal search et beaucoup d'autres moteurs de recherche présente les résultats dans une seule liste classés par rapport à la pertinence de ces résultats à la requête de l'utilisateur (voir figure ci-dessous). Bien que l'on observe une augmentation dans la performance, la qualité des résultats de la recherche mixte n'est pas vraiment évaluée, il n'est pas clair si les résultats retournés sont optimales par rapport aux termes de la recherche agrégée ou pas (i.e. La bonne combinaison des résultats de la recherche web et les résultats verticaux et la bonne position).

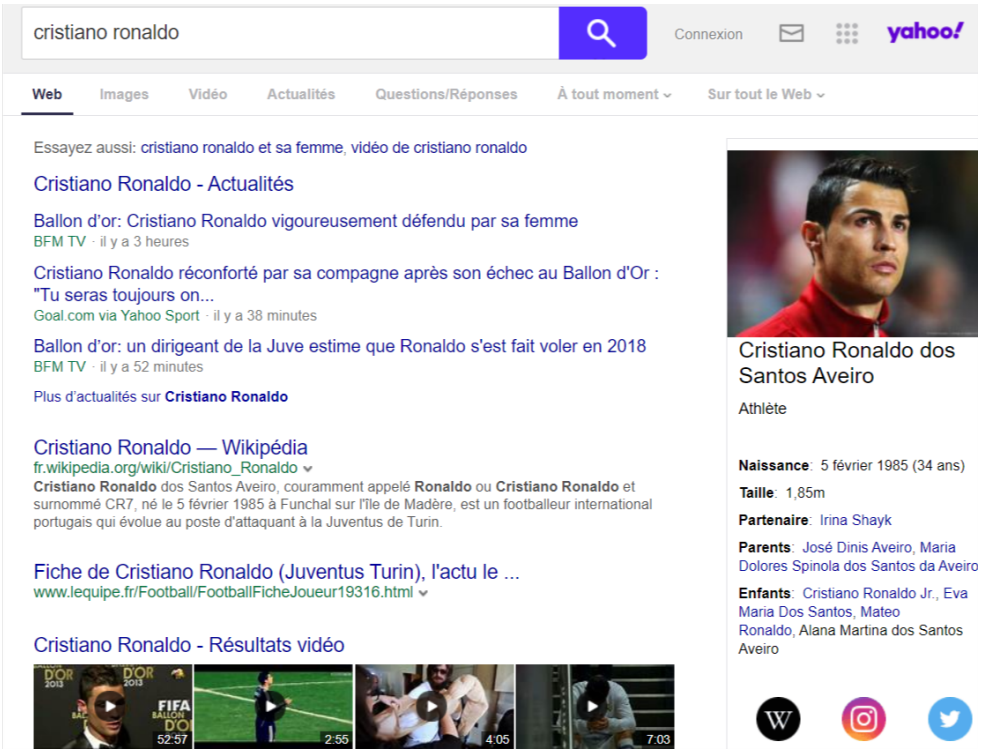


FIGURE 2.6 – Blended Views

Non-Blended views

Dans les vues agrégées non-Blended (non-mixte) les résultats sont regroupés par rapport à leurs type (images, news, vidéos, . . . etc) et ils sont affichés dans des panneaux séparés.

Yahoo Alpha représenté sur la figure ci-dessous est un exemple d'une telle conception, d'autres exemples incluent Kosmix, Naver et Google Searchmash. Quelque soit le minimum de résultats disponible pour une requête, les différents panneaux correspondants sont remplis et affichés. La mise en place des différents panneaux est généralement prédéfinie.

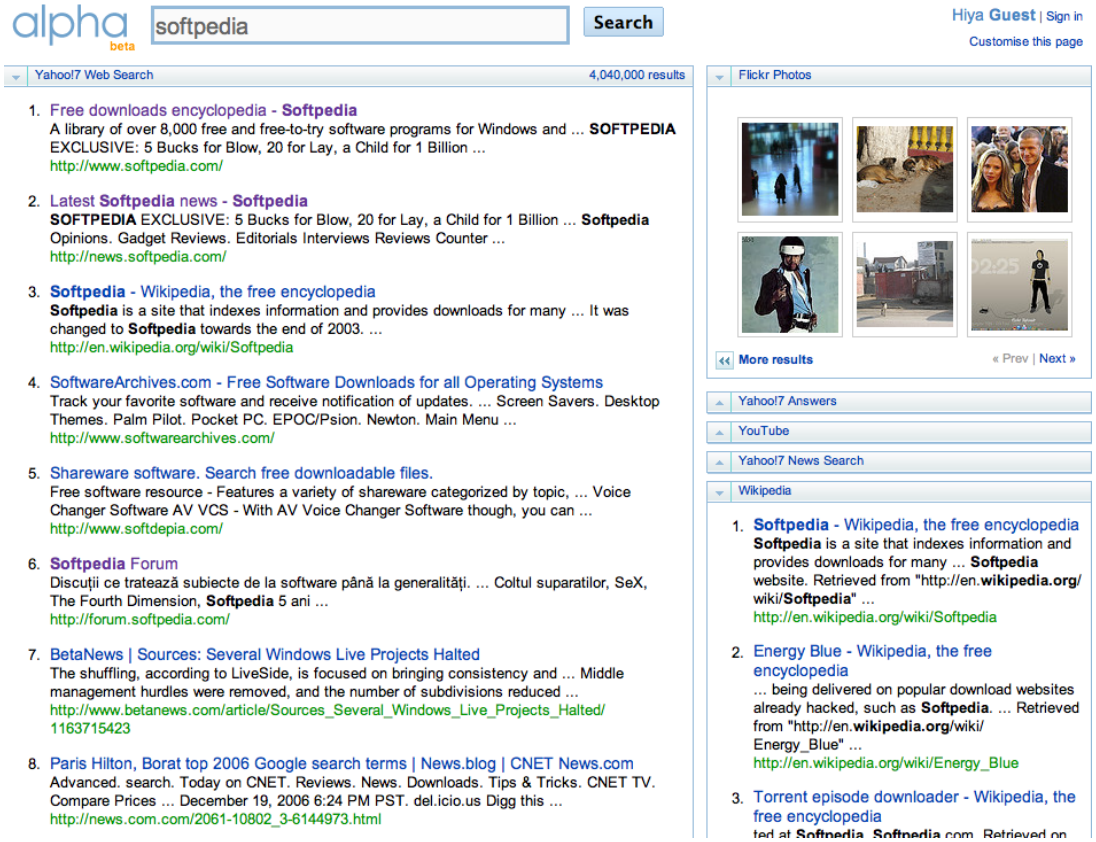


FIGURE 2.7 – Non Banded Views

2.6.5 Agrégation par génération automatique de documents

Les techniques d'agrégation par génération de documents cherchent à construire ou à générer automatiquement un document à partir de plusieurs documents de la même ou de différente sources. Les auteurs dans [Sauper et al., 2009] construisent automatiquement des articles médicaux pour Wikipedia à partir de modèles qu'ils ont eux-mêmes générer. Le contenu est ensuite sélectionné à partir d'Internet pour chaque partie du modèle (le diagnostic, les symptômes, les causes ...). Les auteurs dans [Paris et al., 2009] ont utilisé les techniques de génération du langage naturel (NLG) pour créer un résultat agrégé et cohérent (le système ScjFly de génération de brochures automatiques pour des entreprises). L'organisation de l'information restituée est définie par le rôle que joue chaque unité d'information ainsi que les relations (linguistiques) qui peuvent exister entre les différentes unités d'information restituées

FIGURE 2.8 – Le système ScjFly (génération automatiques de brochures d'entreprises)

2.7 Conclusion

Nous avons donné dans ce chapitre un bref aperçu sur le paradigme de la RI agrégée. Nous avons montré quelques domaines d'application de la RI agrégée ainsi que les problématiques et les approches d'agrégation appliqués dans ce domaine.

Chapitre 3

Système de génération de revue de presse

3.1 Introduction

Après avoir présenté dans les chapitres précédents, un état de l'art sur le paradigme de recherche agrégée ainsi que sur le domaine de revues de presses, nous allons à présent s'intéresser à la conception de notre système. Nous rappelons que l'objectif de notre travail c'est de développer un système de génération automatique de revues de presse. Pour atteindre cet objectif, nous allons considérer ce système comme un cas applicatifs de la RI agrégée. Le processus générique de la RI agrégée, les différents techniques d'agrégation des résultats seront exploités dans le développement de notre système.

3.2 Génération de revue de presses comme un cas d'application de la RI agrégée.

Le choix de se baser sur la RI agrégée comme paradigme sur lequel on s'est basé pour le développement de notre système est justifié par plusieurs considérations.

— Tout d'abord, comme dans le cas de RI agrégée, la génération d'une revue de presse suppose l'interrogation de plusieurs sources d'information (quotidien d'information, sites de news, agence de presse,...).

— Une revue de presse doit contenir des informations variées et hétérogènes (plusieurs point de vues, plusieurs sujets traités dans l'actualité, et plusieurs grille de lecture). La RI agrégée consiste aussi à fournir dans la même page de résultats

des informations variés diversifiées et hétérogènes.

— Enfin et c'est les plus importants éléments, l'agrégation est au centre de tout système de RI agrégée. C'est le cas aussi dans le domaine des revues de presses où la synthèse de toute l'actualité est le but essentiel. La génération automatique d'une revue de presse ne se fait pas aléatoirement, de ce fait on doit suivre une des approches d'agrégation définit précédemment. Il ne s'agit pas seulement de lister tout les articles d'actualité mais de les synthétiser et de les agréger afin d'en faire un résumé de tout les sujets traités.

Partant de ce constat, nous allons donc suivre le processus générique de la RI agrégée (vu précédemment) pour développer notre système. Nous rappelons ici que le processus générique de la RI agrégée est constitué de trois phases[14] :

— La phase dispatching de la requête vers les sources d'information. Cette phase est connue aussi sous le nom de sélection des sources (sources sélection) qui permet de sélectionner pour une requête donnée, les meilleures sources susceptibles de contenir l'information souhaitée.

— La phase de sélection des unités d'information (appelés nuggets) qui consiste a récupérer de chaque sources les informations pertinentes à la requête

— Et enfin, la phase d'agrégation ou de fusion de résultats, qui consiste à organiser les différentes résultats (nuggets) dans la même page de résultats.

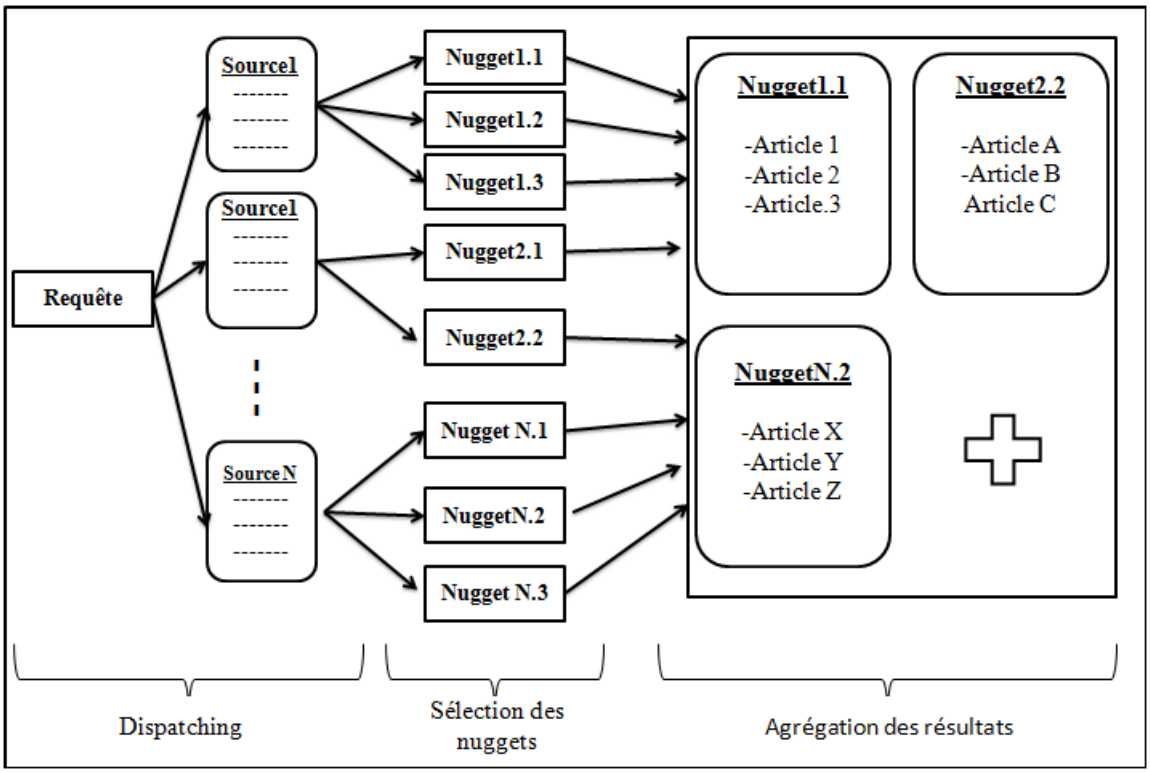


FIGURE 3.1 – Processus de génération de revues de presse

Ces processus est générique, nous devons donc adapter et spécifier chaque étape de ce processus à notre cas. Mais avons cela, nous devons d'abord choisir les sources d'information pour notre système.

3.2.1 Identification des sources

Avant de commencer à élaborer une revue de presse, un journaliste doit d'abord sélectionner (ou choisir) un certain nombre de sources (de journaux) d'où il puisera les différents articles. Dans un système de génération automatique de revue de presses, l'ensemble de sources est sélectionné. Ces sources sont essentiellement des quotidiens de la presse écrite, des sites d'information en ligne, des sites d'agences de presse.... Ce qui est intéressant aussi c'est de donner la possibilité à l'utilisateur de choisir lui-même les sources qui l'intéresse.

Dans le cas de revue de presse thématique (spécialisée), les sources d'information spécialisées dans le thème désiré sont sélectionnées. La sélection peut se faire aussi au niveau des articles en ne sélectionnant que ceux concernés par le thème de la

revue. Par exemple, pour une revue de presses thématique spécialisée dans l'information sportives nationale, les sources suivantes peuvent être choisies.

FIGURE 3.2 – exemple de sources pour une revue de presse « sport national ».

3.2.2 Le dispatching de la requête

La première étape du processus générique de la RI agrégée est le dispatching de la requête. Comme dans tout processus de RI, c'est l'utilisateur qui déclenche le processus en soumettant une requête au système. Pour notre cas, il n'existe pas de requête explicite. Toutes les sources sont utilisées à chaque génération et les articles de news de chacune des sources sont récupérés pour construire la revue de presse.

Toutefois, le système doit donner la possibilité à l'utilisateur de choisir ces propres sources d'informations s'il souhaite personnaliser sa revue de presse. Dans ce cas, seuls les articles des sources choisis par l'utilisateur seront considérés.

Le système doit aussi donner la possibilité à l'utilisateur de définir un thème de la revue à générer. Dans ce cas précis, le thème sera utilisé comme une requête

dans la phase de sélection (phase suivante).

3.2.3 Sélection des nuggets

Dans cette étape, les articles à considérer dans la génération de la revue de presses sont sélectionnés. Cette sélection doit se baser sur les critères suivants :

— Pour une revue de presse généraliste, seul le critère d'actualité (date de publication de l'article) est considéré.

— Dans le cas d'une revue de presses thématique où l'utilisateur choisis un thème particulier pour sa revue, seules les articles en relation avec le thème seront sélectionnés. Bien sur le critère d'actualité reste toujours de mise ici.

— Un critère de sélection thématique est utilisé dans le cas de génération de revue thématique. Dans ce cas précis, le thème (sport, culture, politique...) peut être considéré comme une requête afin de ne retenir que les articles de ce thème.

3.2.4 L'agrégation des résultats

La dernière étape du processus de la RI agrégée consiste en l'assemblage des données récoltées à partir de différents sources.

Le résultat agrégé est une alternative à la traditionnelle liste de documents répondant chacun à une partie du besoin utilisateur. Le principal objectif de l'agrégation est d'organiser les informations d'une façon à faciliter la navigation à l'utilisateur.

Partant du fait qu'une revue de presses est constituée de sujets variés et diversifiés. Notre système doit adapter une approche d'agrégation qui aidera et facilitera à l'utilisateur de :

D'avoir une vue générale sur la liste des sujets (thème) d'actualité traités dans la revue de presse.

D'accéder facilement au sujets qui l'intéresse, d'avoir une synthèse des contenu des articles dans chaque sujet et évidement d'avoir un accès aux articles lié à un sujet donné.

Pour cela, nous proposons de combiner plusieurs techniques d'agrégation comme suit :

— D'utiliser tout d'abord une approche d'agrégation par Clustering en premier lieu permettra d'identifier les différents sujets d'actualité et de regrouper les

articles de news dans un ensemble de cluster correspondant chacun à un sujet traité.

— De combiner ensuite à cette étape de clustering, une techniques de résumé multi-documents. un résumé peut être ajouté pour chacun des clusters construits à partir d'articles de même thématique (même sujet), l'agrégation résumé multi-documents est l'approche plus efficace pour générer ce travaille afin de produire un résumé pour chaque sujets (cluster).

— - Enfin, nous adopterons une technique de vue agrégée de type Blended View pour la présentation des clusters sur la page des résultats.

Le schéma suivant illustre mieux notre explication (figure 3.3).

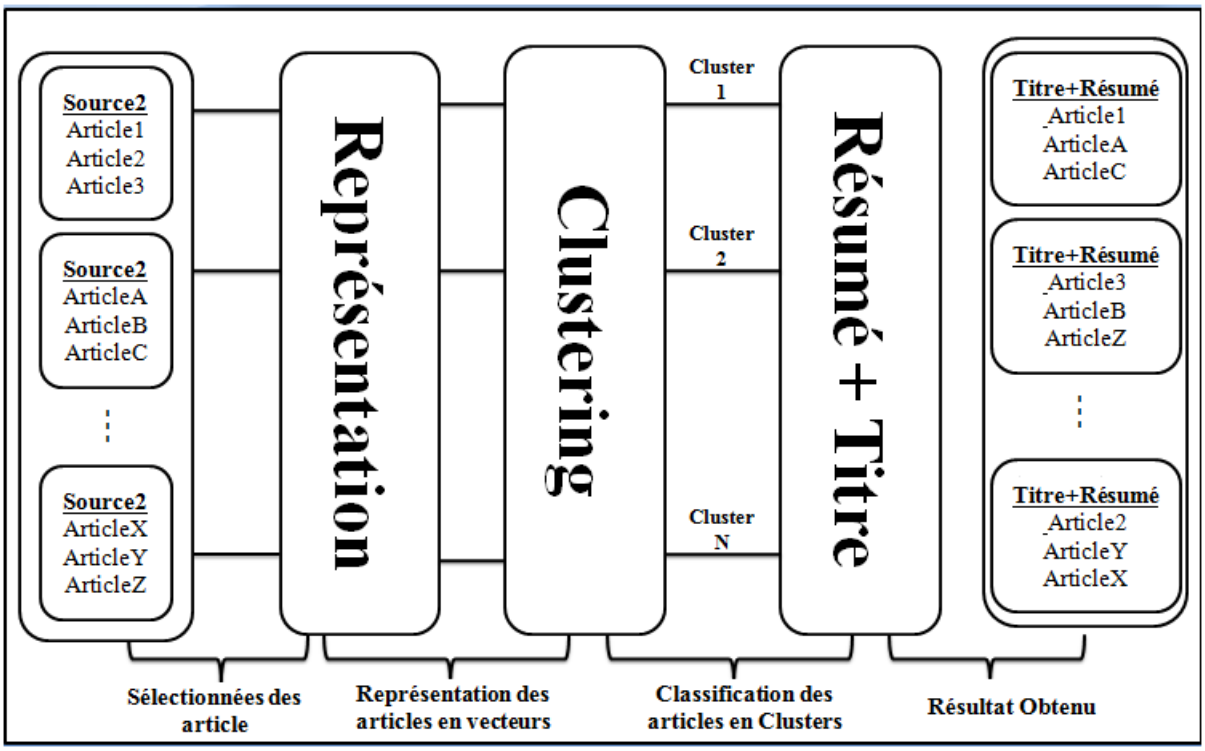


FIGURE 3.3 – La représentation de notre revue

Nous devons maintenant choisir des techniques (ou des algorithmes) pour le Clustering d'articles de news ainsi que pour le résumé automatiques de document qu'on pourra utiliser dans notre système.

3.3 Les étapes de l'aggrégation des resultats

3.3.1 Etape1 : Clustering

L'objectif du Clustering est de former des clusters tels que chaque cluster soit homogène, les éléments de chaque cluster (dans notre cas c'est les articles) doivent être le plus similaires possible, et les clusters doivent être hétérogènes entre eux (deux clusters doivent être le plus dissimilaires possible). La plupart des méthodes utilisent une mesure pour calculer la qualité d'un Clustering en se basant sur une distance intra-cluster et une distance inter-clusters. Les algorithmes cherchent alors à obtenir un Clustering de qualité optimale, correspondant à une distance intra-cluster faible et une distance inter-clusters élevée. Traditionnellement, les techniques de Clustering sont divisées en trois familles principales : les méthodes de partitionnement, les approches hiérarchiques et la classification conceptuelle .[18]
Nous avons opté pour notre système un clustering par les méthodes de partitionnement pour les raisons de :

— Les méthodes de partitionnement sont particulièrement intéressantes dans la construction de clusters/groupes.
— Elles Cherchent la meilleure partition en K classes (clusters ou groupes) disjointes des données.
— Les approches par partitionnement utilisent un processus itératif fonction de nombre K qui consiste à affecter chaque individu à la classe la plus proche au sens d'une distance ou d'un indice de similarité en optimisant une certaine fonction objectif.

Définition du partitionnement : Le partitionnement a pour but l'identification de classes disjointes de documents au sein d'une collection donnée. On cherche bien sûr, d'une part, à obtenir des classes homogènes, c'est-à-dire qui rassemblent des documents proches les uns des autres, et, d'autre part, à éviter que des classes

différentes soient proches l'une de l'autre (si tel était le cas, on aurait intérêt à les fusionner). On dit souvent que l'ensemble des classes structure la collection. Les approches de classification par partitionnement permettent de subdiviser l'ensemble des individus en un certain nombre de classes en employant une stratégie d'optimisation itérative dont le principe général est de générer une partition initiale, puis de chercher à l'améliorer en réattribuant les données d'une classe à l'autre. Il n'est bien entendu pas souhaitable d'énumérer toutes les partitions possibles. Ces algorithmes recherchent donc des maxima locaux en optimisant une fonction objectif traduisant le fait que les individus doivent être similaires au sein d'une même classe, et dissimilaires d'une classe à une autre. Les classes de partition final, prises deux à deux, sont d'intersection vide est représentée par noyau .[19]

Pour obtenir un bon partitionnement, il convient d'à la fois :

— minimiser l'inertie *intra-classe* pour obtenir des grappes (cluster) les plus homogènes possibles.

— maximiser l'inertie *inter-classe* afin d'obtenir des sous-ensembles bien différenciés.

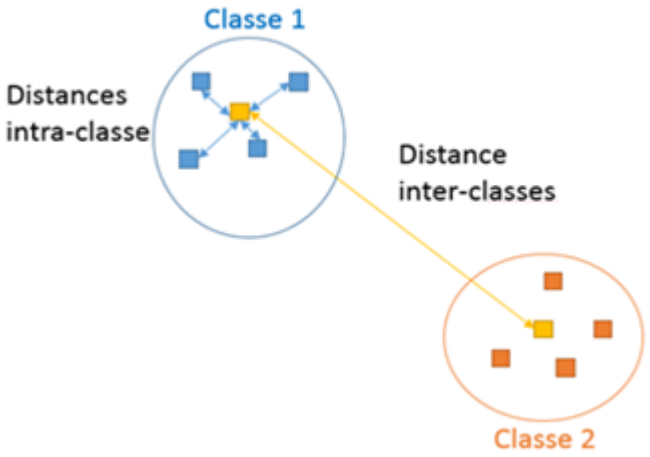


FIGURE 3.4 – Exemple de classification par partitionnement

Le partitionnement d'un ensemble de documents en différentes classes est en général un processus itératif au cours duquel un utilisateur cherche à mieux comprendre le contenu d'une collection, ou à en tirer des informations utiles. Ce processus passe par les étapes suivantes :

— 1 Représentation des documents.
— 2 Choix d'une mesure de similarité et éventuellement calcul de la matrice de similarité.
— 3 Choix de l'algorithme de partitionnement
— 4 Validation des classes obtenues.
— 5 Retour à l'étape 2, en modifiant les paramètres de l'algorithme de partitionnement voire la méthode de partitionnement et l'algorithme associé ; ce retour est bien sûr optionnel.

Représentation des documents (articles) pour le clustering

La représentation la plus simple des textes introduits dans le cadre du modèle d'espace vectoriel est appelé **sac de mots** (en anglais **Bag of Words**). L'idée est de transformer des textes dans des vecteurs où chaque composante représente un mot. Cette représentation des textes exclut toute analyse grammaticale et toute notion de distance entre les mots, et déstructuré syntaxiquement les textes en les rendant compréhensible pour la machin.[20][21]. La plupart des approches sont centrées sur la représentation vectorielle des textes en utilisant la mesure TF x IDF. TF représente « *Term Frequency* » le nombre d'occurrences du terme dans le corpus et IDF représente le nombre de documents contenant le terme. Ces deux concepts sont combinés (Par produit), en vue d'attribuer un plus fort poids aux termes qui apparaissent souvent dans un document et rarement dans l'ensemble du corpus. La formule TF*IDF est la suivant :

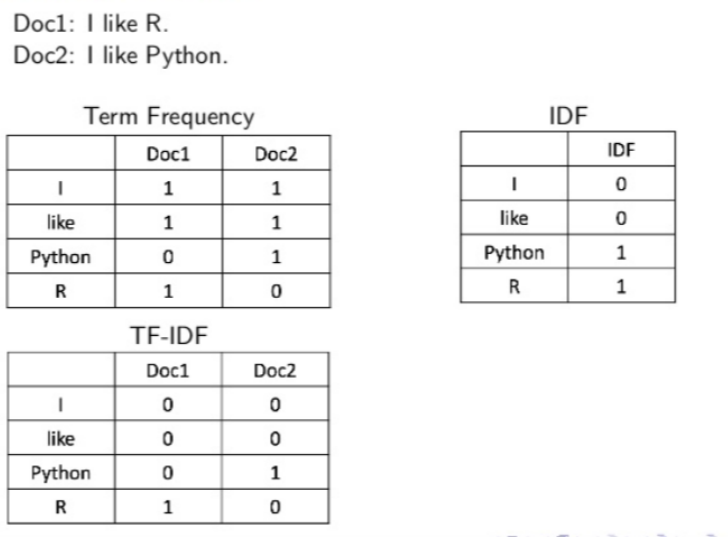
$$\text{TF*IDF}(\text{t}_k, d_j) = Occ(t_k, d_j) * \log \frac{Nb_{doc}}{Nb_{doc}(t_k)}$$

Où :

$\text{Occ}(\text{t}_k, d_j)$ est le nombre d'occurrences (la fréquence) du terme t(k) dans le document d(j) du corpus.

Nb_{doc} est le nombre total des documents du corpus.

$\text{Nb}_{doc}(t_k)$ est le nombre de documents dans lequel le terme t(k) apparaît au moins une fois.


Doc1: I like R.
Doc2: I like Python.

Term Frequency

	Doc1	Doc2
I	1	1
like	1	1
Python	0	1
R	1	0

IDF

	IDF
I	0
like	0
Python	1
R	1

TF-IDF

	Doc1	Doc2
I	0	0
like	0	0
Python	0	1
R	1	0

FIGURE 3.5 – Exemple de représentation de document avec TF*IDF

Choix d'une mesure de similarités

La similarité entre deux éléments i et j est noté s *(i, j)*. La similarité entre les données à étudier est l'information principale (et souvent unique) permettant à l'algorithme de classification de partitionner le jeu de donnée. Le plus souvent, cette similarité est calculée selon une mesure de "distance" adaptée au type de données et au problème à résoudre. Une distance, notée de manière générale *d (i, j)*, est une mesure de similarité qui vérifie certaines propriétés mathématiques (symétrie, séparation et inégalité triangulaire). Il existe bien sûr plusieurs mesures de similarité ou de distance, les plus connues et les plus utilisées étant **l'indice de Jaccard**, **le coefficient de Dice**, **le cosinus et la distance euclidienne**. Toutes ces mesures ont été et sont toujours couramment utilisées dans le cas de données textuelles.[20]

La distance euclidienne : est la plus connue et la plus utilisée pour calculer la distance dans l'espace vectoriel. La distance d (i, j) entre les points xi et xj sous la formule suivante :

$$d(i,j)=\sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + ... + |x_{im} - x_{jm}|^2}$$

Choix de l'algorithme de partitionnement

Il existe nombreux d'algorithmes (méthodes) de partitionnements comme :

— Centres mobiles.

— K –means(Mc Queen).

— Nuées dynamiques (Duday).

— K-représentants (k-médoids).

— Réseaux de kohonen.

— Méthodes basés sur la notion de densité.

Aprés l'étude de toutes ces méthodes, on a décidé que l'approche K –means est la méthode la plus fiable pour notre système (revue de presse).

K –means (k-moyennes)

L'algorithme k-means mis au point par McQueen en 1967 [MacQueen, 1967], un des plus simples algorithmes d'apprentissage non supervisé , appelée algorithme des centres mobiles, il attribue chaque point dans un cluster dont le centre (centroïde) est le plus proche. Le centre est la moyenne de tous les points dans le cluster , ses coordonnées sont la moyenne arithmétique pour chaque dimension séparément de tous les points dans le cluster cad chaque cluster est représentée par son centre de gravité. La somme des dispersions (distances) entre un point et son centroïde, exprimée par une distance appropriée, est utilisée comme fonction objective. L'idée principale est de définir les k centroïdes arbitraires c1, c2,..., ck (k le nombre de clusters fixé a priori, chaque ci représente le centre d'une classe), Ces centroïdes doivent être placés dans des emplacements différents. Donc, le meilleur choix est de les placer le plus possible éloignés les uns des autres. La prochaine étape est de prendre chaque point appartenant à l'ensemble de données et l'associer au plus proche centroïde. Càd Chaque classe Si sera représentée par un ensemble d'individus les plus proches de son ci .[22]

L'algorithme standard associé consiste alors, à partir d'un ensemble de représentants initiaux, à itérer les deux opérations suivantes :

L'algorithme général simple K-Means

Entrées : *k le nombre maximum de classes désiré.*

Début

— 1. Choisir k individus au hasard (comme centre des classes initiales)

— 2. Affecter chaque individu au centre le plus proche

— 3. Recalculer le centre de chacune de ces classes

— 4. Répéter l'étape (2) et (3) jusqu'à stabilité des centres

— 5. Éditer la partition obtenue.

Fin

Algorithme des k-moyennes

Entrée :
– C = {d1, ⋯, dN}, collection de documents ;
– K, nombre de classes ;
– T, nombre d'itérations maximal admis;
Initialisation :
– $(r_1^{(0)}, ..., r_K^{(0)})$, ensemble de représentants initiaux;
– $t \leftarrow 1$;
tant que (t <T) ou (l'ensemble des classes n'est pas stable) **faire**
pour chaque $d \in C$ **faire**
// Etape de réaffectation

$$G_k^{(t)} \leftarrow \left\{ d: \left\| \boldsymbol{d} - \boldsymbol{r_k^{(t-1)}} \right\|_2^2 \leq \left\| \boldsymbol{d} - \boldsymbol{r_l^{(t-1)}} \right\|_2^2, \forall l \neq k, 1 \leq l \leq K \right\};$$

Fin
Pour chaque $k, 1 \leq k \leq K$ **faire**
// Etape de recalcule des centroïdes

$$r_k^{(t)} \leftarrow \frac{1}{|G_k^{(t)}|} \sum_{d \in G_k^{(t)}} \boldsymbol{d};$$

Fin
$t \leftarrow t + 1;$
Fin
Sortie : G, une partition de C en K classes

Validation des classes obtenus

Dans notre cas, le Clustering n'est as une fin en soit, c'est juste une étape importante dans notre processus de génération de brochures touristique. Il est vrai que la qualité du clustering influencera endorment la qualité de notre système.
Pour cette évaluation, nous pouvons opter pour une évaluation avec des métriques d'évaluation connu telle que Le rappel et précision, la F-mesures très connus dans le domaine de la RI ou sinon aller vers une évaluation empiriques avec des utilisateurs du système.

Pour une évaluation avec des métriques, nous devons disposer d'un dataset de test spécifique à notre domaine d'études (articles de news). Ce qui n'est pas le cas. C'est la raison pour laquelle, nous optons pour une évaluation empirique une fois le système seta mis en place.

3.3.2 Etape 2 : Résumé du contenu de chaque cluster

Une technique d'agrégation par résumé multi-document permet de générer des résumés pour un ensemble de documents et portant souvent sur une même thématique (voir le chapitre 'recherche d'information agrégée').
Nous voulons appliquer cette approche pour réaliser un résumé pour chaque groupe (cluster) d'articles similaire (qui parlent sur le même sujet) construit d'après la première étape (Clustering).

3.3.3 Etape 3 :Organisation des résultats (selon la Blended View)

Dans les vues agrégées *Blended View* les résultats de recherche hétérogènes provenant des différentes sources sont fusionnés et présentés verticalement dans une liste. Pour la visualisation de la revue de presse, nous avons opté pour la Blended View ou tout le groupe d'articles construits sera présenté à la mode verticale dans une même interface. figure [h !]

Présentation *'Blended View'*

3.4 Conclusion

Dans ce chapitre, nous avons présenter le fonctionnement générale de notre système de génération de revue de presse. Pour cela nous nous somme basé sur le paradigme de la RI agrégée qui nous a fourni le principe, les techniques et le cadre pour la réalisation de notre système. Nous avons aussi présenté nos choix de méthodes et d'algorithmes pour la phase de Clustering des articles de news qui est au centre de notre système. Dans le prochain chapitre, nous présentons l'application que nous avons développer pour matérialiser tout nos choix de conception.

Chapitre 4

L'implémentation

4.1 Introduction

Après avoir finalisé l'étape de conception, nous consacrons ce chapitre à la réalisation. Les différentes problématiques ont été profondément analysées, ce qui nous a permis d'entreprendre le développement de notre solution, ayant comme objectif d'aboutir à un produit final exploitable.

Nous allons d'abord présenter l'environnement de travail ainsi que les outils et les logiciels utilisés. Et pour expliquer le fonctionnement de notre prototype on a décrit les interfaces offertes par l'application en utilisant les captures d'écran.

4.2 Environnement et outils de travaille

4.2.1 Les Postes de travaille

Poste de travaille 1 :

Système d'exploitation	Windows 10 Professionnel
RAM	16 Go
Processeur	Intel Core i7-4700MQ CPU @ 2.4GHz
Type de système	Système d'exploitation 64 bits

Table 4.1 – Caractéristiques du poste de travail 1

Poste de travaille 2 :

Système d'exploitation	Windows 7 Professionnel
RAM	6 Go
Processeur	Intel Core i3-4005MQ CPU @ 1.7GHz
Type de système	Système d'exploitation 64 bits

TABLE 4.2 – Caractéristiques du poste de travail 2

4.2.2 Langages de programmation

Nous avons utilisé au cours de la réalisation de notre système, plusieurs langages de programmation et logiciels. Ci-après une brève présentation de ces derniers.

Le langage de script PHP :

(HypertextPreprocessor) : langage de programmation contenu dans des pages Web et exécuté sur les serveurs, ils renvoient directement le résultat vers le client qui ne peut jamais voir le code source. Permet de créer des pages Web dynamiques [23].

JavaScript :

Langage de programmation de scripts principalement employé dans les pages web interactives, mais aussi pour les serveurs avec l'utilisation (par exemple) de Node.js. Il supporte le paradigme objet, impératif et fonctionnel. JavaScript est le langage possédant le plus large écosystème grâce à son gestionnaire de dépendances npm, avec environ 500 000 paquets en août 2017.

HTML :

(HyperText Mark up Language) : est le format de données conçu pour représenter les pages web. C'est un langage de balisage permettant d'écrire de l'hypertexte, d'où son nom. HTML permet également de structurer sémantiquement et de mettre en forme le contenu des pages, d'inclure des ressources multimédias dont des images, des formulaires de saisie, et des programmes informatiques [23].

CSS :

(Cascading Style Sheets) : est un langage déclaratif simple pour mettre en forme des pages HTML ou des documents XML. Le langage CSS permet de préciser les caractéristiques visuelles et sonores de présentation d'une page Web [23].

JQUERY :

C'est un Framework développé en JavaScript qui permet notamment de manipuler aisément la DOM, d'utiliser AJAX, de créer des animations... La vocation première de ce Framework est de gagner du temps dans le développement des applications [24].

JSON

: (JavaScript Object Notation) est un format d'échange de données léger. Il est facile pour les humains de lire et d'écrire. Il est facile pour les machines d'analyser et de générer. JSON est un format de texte totalement indépendant du langage mais qui utilise des conventions bien connues des programmeurs de la famille des langages C, notamment C, C ++, C , Java, JavaScript, Perl, Python et bien d'autres. Ces propriétés font de JSON un langage d'échange de données idéal [25].

4.2.3 Logiciels et éditeurs de textes

AppServ

Appserv est un outil OpenSource pour de Windows avec Apache, MySQL, PHP et d'autres ajouts, où ces applications sont configurées automatiquement, vous permettant d'exécuter un serveur web complet. Comme fonctionnalités supplémentaires phpMyAdmin pour gérer MySQL facilement une plateforme permettant l'exploitation d'un site web en PHP qui éventuellement aurait besoin d'un accès à une base de données [26].

LaTeX

est un langage et un système de composition de documents. Il s'agit d'une collection de macro-commandes. LaTeX permet de rédiger des documents dont la mise en page est réalisée automatiquement en se conformant du mieux possible à des normes typographiques. Nous avons utilisé LaTeX pour la rédaction de notre mémoire.

Visual Studio Code

(VSC) est un éditeur de code open-source, gratuit et multi-plateforme (Windows, Mac et Linux), développé par Microsoft, à ne pas confondre avec Visual Studio, l'IDE propriétaire de Microsoft. VSC est développé avec Electron et exploite des fonctionnalités d'édition avancées du projet Monaco Editor. Principalement conçu pour le développement d'application avec JavaScript, TypeScript et Node.js, l'éditeur peut s'adapter à d'autres types de langages grâce à un système d'extension bien fourni [27].

Sublime Texte

Sublime Texte : Éditeur de texte générique codé en C++ et Python, il est disponible sur Linux, Mac et Windows. Depuis la version 2.0, sortie le 26 juin 2012, l'éditeur prend en charge 44 langages de programmation majeurs, tandis que des plugins sont souvent disponibles pour les langages plus rares. Nous avons utilisé la version d'essai qui est en libre accès sur internet.

APIs utilisé

NewsApi : Web API qui permet d'obtenir des articles de presse de dernière minute et de rechercher des articles de plus de 30 000 sources et blogs[28]. Elle fournit également la possibilité de sélectionner les sources, les pays, les catégories, etc.
On a utilisé cette API car :

— Il est difficile d'avoir un accès aux sources d'information en local.
— Il nous permet d'accéder à plus de 30 000 source d'information.
— Il nous retourne des articles avec les méta-donnée (Source,Titre,URL,Contenu,...).

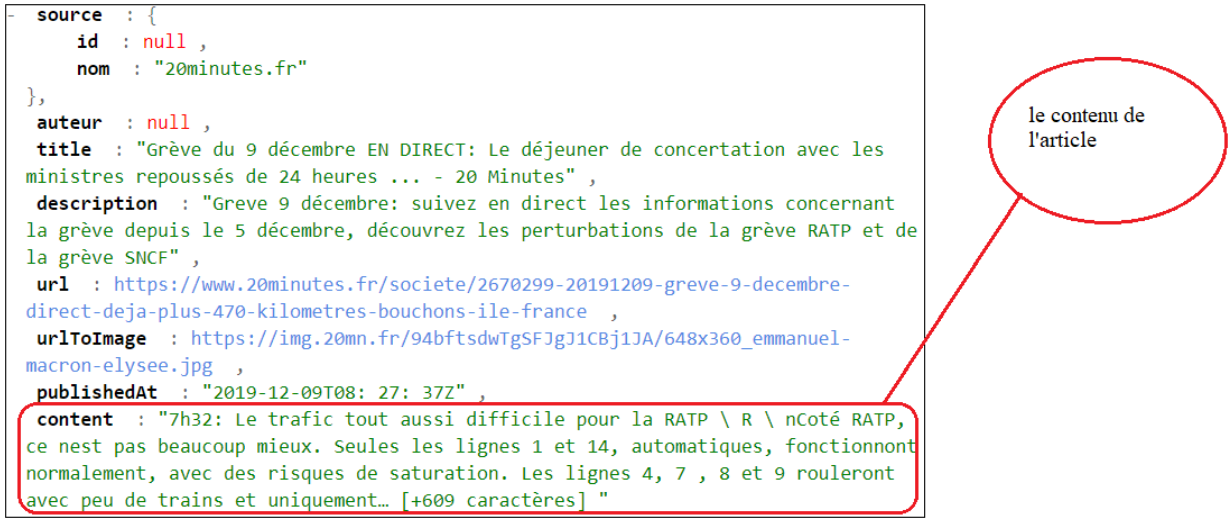


FIGURE 4.1 – Exemple d'un article retourné(Format JASON)

4.3 Application

4.3.1 Page d'Accueil

C'est la page principale de site où on trouve les articles récents récupérés des plusieurs sources différentes .

FIGURE 4.2 – Page d'Accueil de site

4.3.2 Articles par Catégories

Catégories Disponibles :

les utilisateurs peuvent consulter et y accéder aux catégories disponibles à partir d'un panneau latéral.

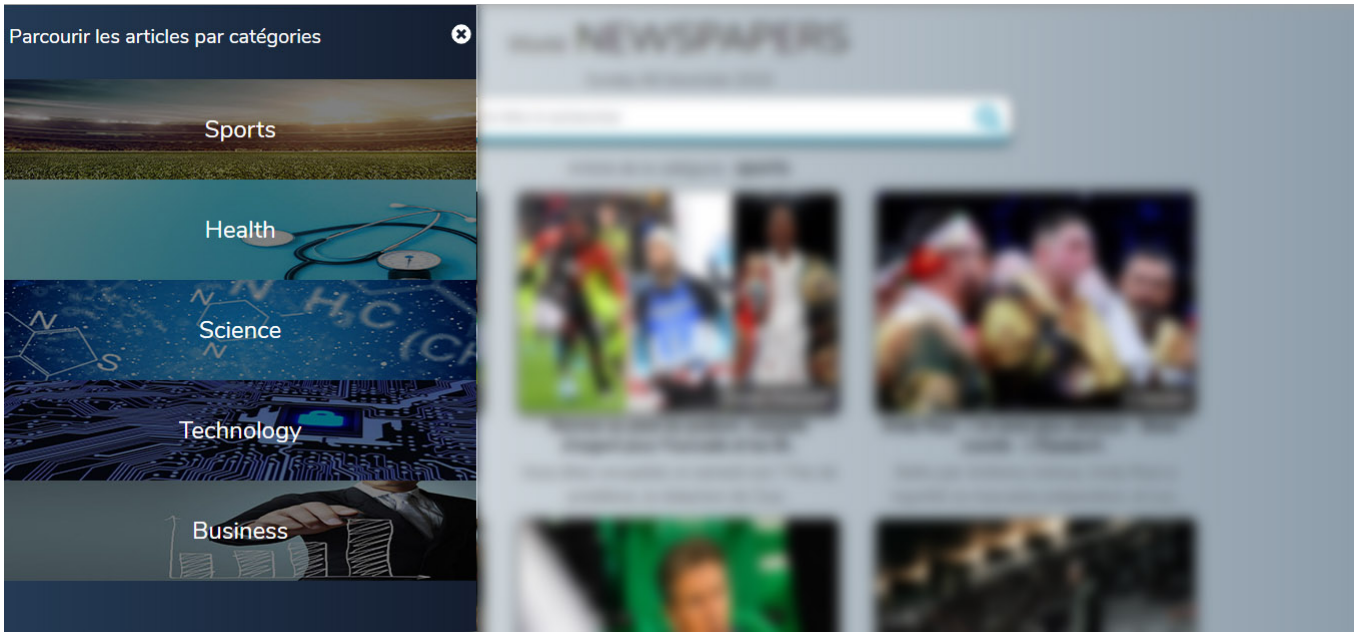


FIGURE 4.3 – Les catégories disponibles

Articles regroupé par catégories :

Une fois l'utilisateur choisis la catégorie à consulter, le système l'oriente vers une autre page qui contient seulement des articles de cette catégorie

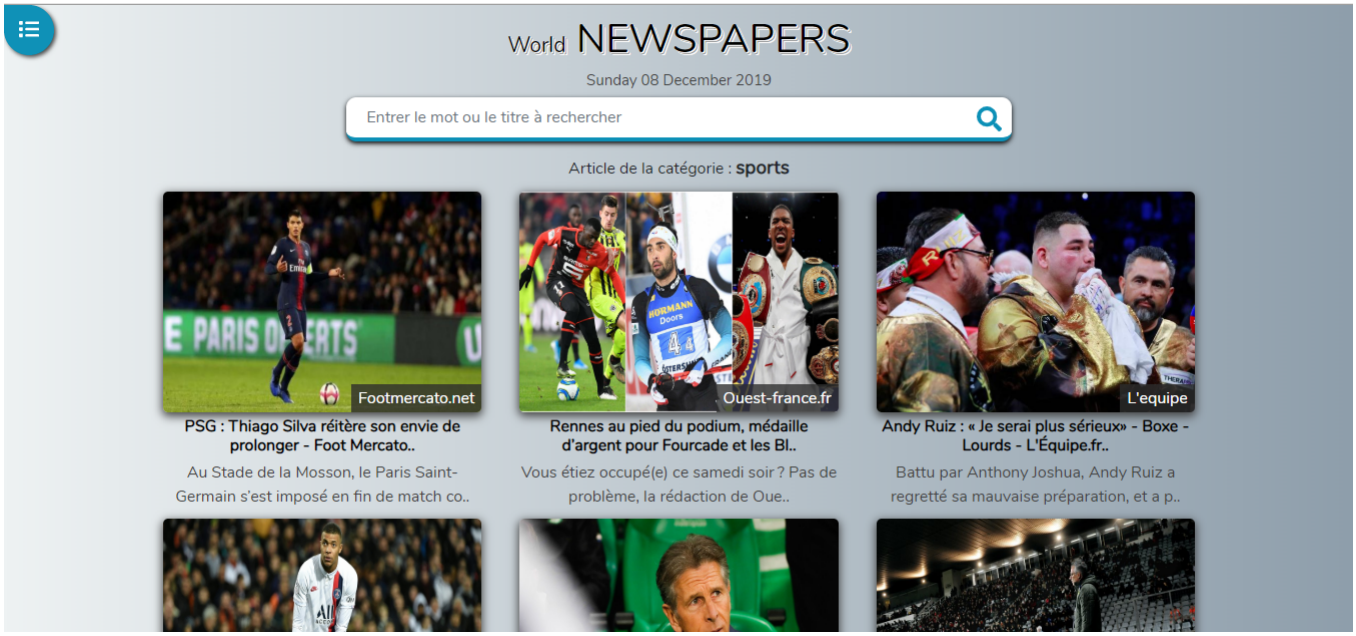


FIGURE 4.4 – Articles regroupé par catégories

4.3.3 Recherche des Articles :

La plateforme offre aux utilisateurs un champs de recherche, après la validation de ce dernier l'utilisateur sera orienté vers une autre page contenant les articles conformes au terme recherché

FIGURE 4.5 – Recherche des Articles :

4.4 Conclusion

Dans ce chapitre, on a présenté les différents outils et langage utilisés dans l'implémentation de notre application, ainsi que la présentation de notre application en utilisant quelques interfaces avec des fonctions qui sont en mesure de satisfaire les besoins des utilisateurs du système.

Conclusion générale

Le travail qui nous a été demandé dans le cadre de notre projet de mémoire de Master consistait à développer un système de génération automatique de revues de presse. Il est connu qu'aujourd'hui, les internautes se retrouvent quotidiennement face à un flux d'informations massif et une diversité de contenu très importante ; les médias et sites d'actualité y contribuent pleinement. Dans un environnement où l'utilisateur ne dispose toujours pas de moyen satisfaisant pour gérer ce flux, un nouveau besoin se crée, celui d'une expérience de lecture synthétisée, personnalisée, correspondant le plus pertinemment et efficacement aux besoins et centres d'intérêt de l'utilisateur.

Les travaux présentés dans ce mémoire se situent dans le contexte de la recherche d'information, et plus particulièrement dans le cadre de la recherche d'information agrégée. Les systèmes de recherche d'information classiques renvoient à l'utilisateur une liste ordonnée répondant au mieux à la requête. Ce paradigme s'est vu complété voire modifié, par la Recherche d'Information agrégée pour répondre au problème de l'existence de plusieurs éléments pertinents dans le même document. La recherche agrégée propose d'assembler dans un seul document des contenus pertinents, complémentaires provenant de sources différentes.

Nous nous somme basé sur ce paradigme pour le développement de notre système. Cela est justifié par le grand nombre de similitudes existant entre ce paradigme et le notion de génération de revues de presse. Nous avons pour cela adopté le processus générique de la RI agrégée présenté par **Arlind Kopliku** (Arlind Kopliku, 2009), et nous l'avons adapté à notre cas d'études.

Notre objectif, dans ce travail, est de concevoir un mini agrégateur d'articles de presse en étudiant et exploitant les approches d'agrégation. Les articles renvoyées par le mini agrégateur seront extraits et sélectionnés a partir de différentes sources d'actualités qui seront ensuite organisés selon les différentes approches d'agrégations étudiées.

Ce projet nous a été bénéfique sur plusieurs plans. Il nous a permis de mettre en pratique toutes les connaissances acquises lors de notre cursus de Master en Systèmes d'Information et Logiciels. Le mémoire nous a permis d'améliorer notre capacité rédactionnelle et de devenir plus rigoureux quant aux questions scientifique, des sources d'informations et de leur référencement. Le travail en binôme, supervisé par notre promoteur a renforcé notre capacité à travailler en équipe.

Nous avons tenté d'implémenter une application simple et efficace en exploitant des ressources existantes sur le web, en exploitant différentes api de site web, d'autres moteurs de recherche ou des bases de données. Cependant on a rencontré énormément de difficultés lors de la conception et de l'implémentation.
Toutefois, et bien qu'étant arrivés à un résultat satisfaisant, il existe quelques points qui nécessitent d'être améliorer. Nous pouvons citer à titre d'exemple :

— Augmenter le nombre de sources d'articles qui alimentent le système pour couvrir plus de sujets et thèmes.

Bibliographie

a. Bibliographie :

-[1] F. MARTIAL, collège de Navarre, Evreux

-[2] Moureau (F.), Répertoire des nouvelles à la main : dictionnaire de la presse manuscrite clandestine, XVIe-XVIIIe siècles, Oxford, Voltaire Foundation, 1999.

-[3] Gunter Volz (coord.), Individu et autorités : positions de la Presse des Lumières, CRINI, Nantes, 2004, 353 pages.

-[4] G. Clémendot. CDI du Lycée Pierre Larousse de Toucy (89). ECJS 1ère L – Année 2013-2014

-[5] Miles, Alisha (2009). "RIP RSS : relancer des programmes innovants grâce à des services vraiment avisés". Journal of Hospital Librarianship.

-[6] Doree, Jim (2007-01-01). Le journal de thérapie manuelle et manipulatrice.

-[7] Serge Courrier, Produire des fils RSS, Editions de l'ADBS, 2009

-[9] Noam Cohen, « Google Actualité », The New York Times, 21 juin 2009

-[11] Google à deux cent pour cent, De Tara Calishain, Rael Dornfest, 2011

-[12] Hervé Didier, « Google Actualités se décline en version audio » (consulté le15 janvier 2019).

-[13] (eds), SIGIR Workshop on Aggregated Search. ACM. 2008.

-[14] Arlind Kopliku. "Aggregated search : From information nuggets to aggregated documents". In : CORIA 09 RJCRI «Rencontre Jeunes Chercheurs en Recherche d'Information ». Toulon, France, 2009.

-[15] K. Arlind, Boughanem Mohand et Collision Karen S. Technical report-IRIT : Aggregated search : potential, issues and evaluation. 2009.

-[16] J. Carbonell et J. Goldstein. "The use of mmr, diversity-based re-ranking for reordering documents and producing summaries". In : In Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval. 1998, 335–336.

-[17] K Ines, B Mohand K Arlind et P Karen. "Une approche de recherche d'attributs pertinent pour l'agrégation d'information". Thèse de doct. Université de Toulouse, 2011.ter

-[18] K Ines, B Mohand K Arlind et P Karen. "Une approche de recherche d'attributs pertinent pour l'agrégation d'information". Thèse de doct. Université de Toulouse, 2011.

-[19]Losee, R.M. (1998). Text retrieval and filtering analytic models of performance. Kluwer Academic Publishers.

-[20] Abdelmalek Amine, Zakaria Elberrichi, Michel Simonet : Evaluation of text clustering methods using wordnet. Int. Arab J. Inf. Technol. 7(4) : 349-357 (2010). http ://www.ccis2k.org/ia

-[21]Mathieu Stricker, 2000 « Réseaux de neurones pour le traitement automatique du langage : conception et réalisatiion de filters d'informations ». Thèse de Doctorat de

l'Université Pierre et Marie Curie URL : http ://www.neurones.espci.fr/

-[22][MacQueen, 1967] : J. B. MacQueen (1967) : "Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability", Berkeley, University of California Press,no 1, pp281-297.1967.

---

b. Webographie :

-[8] Google News : Définition - Louis Maîtreau, 01/10/2019.Disponible sur : https ://www.louismaitreau.fr/definition/google-news/

-[10] https ://news.google.fr

-[23][En ligne]. Disponible sur : http ://dictionnaire.phpmyvisites.net/definition-PHP-4899.html

-[24] Réalisation d'un site web de vente en ligne Taouli Sarah Hamza Cherif Ikram promotion 2014/2015 Université Abou Bakr Belkaid-Tlemcen

-[25] https ://www.json.org/

-[26] https ://www.appserv.org/.

-[27] https ://edutechwiki.unige.ch/fr/Visual$_s tudio_c ode$.

-[28] https ://newsapi.org/.

Résumé

Avec l'émergence des technologies de l'information et de la communication et la démocratisation d'Internet, le volume d'information ne cesse d'accroitre du jour en jour. Dans le domaine de la presse et des média, les sources d'informations se sont multipliées et diversifiées. Nous sommes inondés de news et d'actualité tout au long de la journée et même la nuit. Il devient très difficile pour un citoyen ou une entreprise de suivre tout ce flux d'actualité et de trier les sujets qui l'intéresse. La revue de presse qui est une synthèse des sujets d'actualité traités par les différentes sources d'actualité devient un outil crucial et très utile dans cette situation.

Nous avons développé dans le cadre de ce projet, un système de génération automatique de revues de presse. Notre système collecte les articles de news à partir d'un ensemble de sources d'actualités, classe les articles selon les différents sujets traités et les présente sous forme d'une revue de presse organisé et facile à exploiter.

Mots clés :Presse, Revue de presse, recherche d'information agrégée, Clustering, agrégation de contenu.. . .

Abstract

With the emergence of information and communication technologies and the democratization of the Internet, the volume of information continues to grow from day to day. In the field of news and media, the information sources have multiplied and diversified. We are inundated with news and current affairs throughout the day and even at night. It becomes very difficult for a end user or a company to follow all this news flow and to sort the topics of interest. The press review, which is a synthesis of news topics invoqued by the different news sources, becomes a crucial and very useful tool in this situation.

As part of this project, we developed a system for the automatic generation of press reviews. Our system collects news articles from a variety of news sources, categorizes articles according to different topics and presents them as an organized and easy-to-use press review.

Key words : Press review, aggregated search, Clustering, Content aggregation . . .