Cadre de thèse de doctorat pour la Chaire « Smart Intelligence » de l'ESILV

Les outils de veille technologique et stratégique permettent de délivrer des services de recherches d'information et de notifications de données ciblées, que ce soit en direct ou en temps réel. Ces données ciblées correspondent à des évolutions technologiques visibles sur le Web pour lequel un expert du domaine souhaite rester au courant de la concurrence ou des usages dans son périmètre.

La difficulté pour ces outils de veille est de devoir traiter d'une part les données avec à la fois la multitude de domaines d'expertise pour répondre à la demande des experts, acquérir et gérer un grand volume de données à récupérer sur le Web, analyser le contenu des informations pour en ressortir de la pertinence. Et d'autre part, gérer le profil des experts sur leurs usages de recherche, d'interactions avec la plateforme de veille, mais également les connaissances de l'expert sur son environnement, comme sa propre base de connaissances ou un réseau d'experts.

La société Coexel se positionne dans ce domaine de la veille technologique & stratégique en proposant la plateforme *MyTwip* dédié à ne nombreux domaines d'expertise, avec un moteur de recherche dédié, intégrant des analyses sémantiques basées sur une ontologie pour classifier automatiquement les informations par domaines d'expertise, du traitement de textes pour identifier des signaux faibles pour détecter ces évolutions technologiques pertinentes, ou de l'extraction de connaissances pour relier les concepts liés à une information.

Afin de mieux intégrer l'expert dans le processus de veille, nous envisageons d'intégrer le profil utilisateur, l'expert, au sein même de l'environnement de recherche à différents niveaux. En effet, en intégrant ses recherches ainsi que les interactions effectuées, l'intégration de ses connaissances, un réseau social reliant les experts par affinités de domaines, le tout pouvant produire des informations en temps réel, cela donne une dimension complexe à l'environnement d'analyse et de définition de la pertinence. En effet, il est nécessaire de se focaliser sur la notion de distance entre un expert et la donnée ciblée. Ce profil a pour conséquence de redéfinir cette distance pour l'adapter à l'utilisateur et permettre à l'expert de recevoir des informations plus pertinentes.

La complexité de cette approche réside dans la combinaison de critères :

- Le système doit traiter la donnée à la fois à la demande (moteur de recherche) et en temps-réel : il est donc nécessaire de délivrer une information dans une architecture Lambda (Marz et Warren 2015) tout en respectant la pertinence. Dans cet environnement, le *Batch Layer* stocke l'ensemble des données utiles (information et profil) permettant d'effectuer des recherches pertinentes à la demande, tandis que le *Speed Layer* doit maximiser le profil utilisateur pour traiter le flux de données en temps réel pour générer des notifications pertinentes.

- La pertinence d'une information, que ce soit dans la *Batch* ou *Speed Layer*, repose sur la combinaison subtile de plusieurs critères : des recherches étendues (sémantique, taxonomies, proximité, etc.) et des profils des experts complexes (historiques, réseau, etc.). Cela nécessite la définition d'une distance adaptée entre la donnée et la requête utilisateur, rentrant dans le cadre de la réécriture de requêtes (He, et al. 2016, Grbovic, et al. 2015), afin de produire des résultats pertinents à l'utilisateur. Le profil des experts repose sur plusieurs dimensions : l'historique des données précédemment lues/étiquetées/consultées (données explicites vs implicites), un réseau social d'experts impliquant une propagation de l'information basé sur la proximité d'intérêt, un graphe de connaissances dédié par expert regroupant les informations proches pour représenter les besoins de l'expert sous formes de « concepts » (Wang, Tan et Zhang 2010, Grossetti, et al. 2018). Il est à noter que ces dimensions peuvent être incomplètes, impliquant une adaptabilité pour la réécriture de requêtes. De plus, les

usages des experts évoluant au cours du temps, la pertinence des résultats peut se dégrader. Il est nécessaire de rendre ces mesures auto-adaptatives pour leur permettre améliorer la qualité des résultats.

- Les notifications produites par la *Speed Layer*, doivent être traité en temps réel et prendre en compte plusieurs critères : la temporalité de l'information (un expert s'intéresse aux données récentes), la nouveauté de l'information (pas de redondance), la mutualisation des recherches de nombreux experts ayant souscrits au système. Le domaine du *Publish/Subscribe* répond à ce besoin en optimisant en temps réel les recherches par pertinence et nouveauté (Travers et du Mouza 2018). Les systèmes de recommandations répondent également à la question tout en peinant à répondre au problème de la temporalité des données (Ludmann 2015, Siddiqui, et al. 2014, Subbian, Aggarwal et Hegde 2016).

Ainsi, la croisée de ces différents critères produit un système complexe dont le mélange particulier a pour but de produire des informations pertinentes aussi bien par recherche à la volée qu'en temps-réel. Cette combinaison subtile n'est pas traitée dans la littérature, en effet nous comptons pouvoir produire des recommandations pertinentes de manière efficace en temps réel avec des profils multidimensionnels en se basant à la fois sur des historiques d'événements et sur des graphes de connaissances ou un réseau social.

Ce défi à relever repose sur certains verrous que nous devons soulever :

- Définir une mesure de pertinence de recommandations reposant sur un profil utilisateur riche, reposant sur son historique étendu (données explicites & implicites), son réseau social et son graphe de connaissances ;

- Définir un système optimisé pour la recherche d'information et la recommandation de veille technologique, combinant temps-réel et traitements lourds pour des milliers d'expert.

Ainsi, ce travail de recherche nécessite une thèse de doctorat reposant sur des connaissances en : bases de données (de préférence *Continuous Databases*), recherche d'information, théorie des graphes, web sémantique.

La thèse sera financée par un contrat CIFRE avec Coexel, en partenariat avec le laboratoire DVRC de l'Association Léonard de Vinci (Paris la Défense) au sein du groupe digital, encadrée par Nicolas Travers (HDR).

Bibliographie

Grbovic, Mihajlo, Nemanja Djuric, Vladan Radosavljevic, Fabrizio Silvestri, et Narayan Bhamidipati. 2015. «Context- and Content-aware Embeddings for Query Rewriting in Sponsored Search.» *(SIGIR'15) Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval.* New York: ACM. 383--392.

Grossetti, Quentin, Camélia Constantin, Cédric du Mouza, et Nicolas Travers. 2018. «An Homophily-based Approach for Fast Post Recommendation in Microblogging Systems.» *(EDBT'18) Proceedings of the 21th International Conference on Extending Database Technology.* Vienna: OpenProceedings.org.

He, Yunlong, Jiliang Tang, Hua Ouyang, Changsung Kang, Dawei Yin, et Yi Chang. 2016. «Learning to Rewrite Queries.» *(CIKM'16) Proceedings of the 25th ACM International on Conference on Information and Knowledge Management.* New York: ACM. 1443--1452.

Ludmann, Cornelius A. 2015. «Online Recommender Systems Based on Data Stream Management Systems.» *(RecSys '15) Proceedings of the 9th ACM Conference on Recommender Systems.* Vienna: ACM.

Marz, Nathan, et James Warren. 2015. *Big Data: Principles and best practices of scalable realtime data systems.* Greenwich, CT: Manning Publications Co.

Siddiqui, Zaigham Faraz, Eleftherios Tiakas, Panagiotis Symeonidis, Myra Spiliopoulou, et Yannis Manolopoulos. 2014. «xStreams: Recommending Items to Users with Time-evolving Preferences.» *(WIMS '14) Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS14).* New York: ACM.

Subbian, Karthik, Charu Aggarwal, et Kshiteesh Hegde. 2016. «Recommendations For Streaming Data.» *(CIKM'16) Proceedings of the 25th ACM International on Conference on Information and Knowledge Management.* New York: ACM. 2185--2190.

Travers, Nicolas, et Cédric du Mouza. 2018. «Relevant Filtering in a Distributed Content-based Publish/Subscribe System.» Dans *NoSQL Data Models - Trends and Challenges*, de Olivier Pivert, 193--226. John Wiley & Sons.

Wang, Ziqi, Yuwei Tan, et Ming Zhang. 2010. «Graph-Based Recommendation on Social Networks.» *(APWeb'10) Advances in Web Technologies and Applications.* Busan, Korea: IEEE. 116--122.