Nomao : la recherche géolocalisée personnalisée

Laurent Candillier

Wikio Group
laurent@nomao.com
http://www.nomao.com

1 Introduction

Nomao est un moteur de recherche de lieux qui classe les résultats en fonction de vos goûts et ceux de vos amis. Son développement soulève de nombreuses problématiques scientifiques : extraction et structuration de contenu local, interprétation de requête et recherche d'information, classement de résultats, personnalisations et recommandations.

Présent sur le web et en application iphone et android, Nomao permet à ses utilisateurs de chercher des lieux (restaurants, bars, hôtels, magasins, hôpitaux, musées, etc.), par défaut dans leur environnement proche (en fonction de leur position GPS sur téléphone mobile, s'ils ont accepté de la donner).

Nomao compte aujourd'hui environ 3 millions de lieux référencés, autant de visiteurs uniques par mois, et il répond aux requêtes en 200 millisecondes en moyenne. Outre les aspects techniques, de nombreuses problématiques d'extraction et de gestion de connaissances sont soulevées pour obtenir un tel produit.

La première étape du processus consiste à récupérer un maximum de déclarations de lieux, puis à agréger les différentes informations obtenues pour retourner une fiche unique et aussi complète que possible pour chaque lieu.

À partir d'une telle base de données, les utilisateurs peuvent alors formuler des requêtes pour trouver les lieux qui les intéressent. Il faut alors présenter les lieux répondant aux requêtes, en les ordonnant en fonction d'une certaine notion de pertinence.

Mais contrairement à Google Maps, Nomao va plus loin pour personnaliser les résultats des utilisateurs prêts à fournir davantage d'informations sur leurs attentes. S'ils précisent les lieux qu'ils aiment, le moteur de recherche peut affiner ses réponses en fonction de ces goûts. Et s'ils relient leur compte Nomao à leur compte Facebook, alors les goûts de leurs amis peuvent aussi être utilisés.

2 Extraction et structuration de contenu

Qui dit *Recherche d'Information* dit avant tout information : aussi complète et structurée que possible. La première problématique rencontrée par un moteur de recherche comme Nomao est donc l'acquisition et le traitement des données de lieux qu'il vise.

Chacun des lieux recueillis est automatiquement associé à une ou plusieurs catégories pré-établies : manger, bouger, visiter, etc. Avant même cet automatisme, l'établissement de ces catégories et des mots clés associés est déjà une problématique en soi.

Afin d'enrichir la description de chaque lieu, les notations et commentaires associés à ces lieux sont également récupérés. On entre alors dans le champ du *Traitement Automatique de la Langue*. Une méthode d'extraction des termes présents dans les textes est appliquée. Dans le cadre de la *fouille de données d'opinion*, une base de données d'étiquettes positives et négatives associées à de nombreux termes permet de saisir la tonalité d'un commentaire.

Au final, la base de données de lieux de Nomao contient principalement les informations suivantes : nom du lieu, géolocalisation, adresse, numéro de téléphone, site web, ensemble de catégories / mots clés, ensemble de notations / commentaires.

Mais à ce niveau, un même lieu peut apparaitre plusieurs fois en base, car il a été récupéré depuis différentes sources. Il faut donc identifier ces doublons et les agréger. On aborde ici la problématique de la *déduplication* [Sarawagi et Bhamidipaty (2002)].

À partir d'une base de données de couples de lieux étiquetés positivement (s'il s'agit du même lieu) ou négativement (s'il s'agit de lieux différents), un algorithme d'*apprentissage supervisé* peut être utilisé pour prédire si deux déclarations font référence au même lieu.

Les attributs associés aux exemples d'apprentissage concernent la comparaison des caractéristiques des couples de lieux considérés : similarités entre les noms, géolocalisations, adresses, téléphones, sites web, catégories et mots clés. Les mesures de similarité classiques entre chaînes de caractères sont adaptées : Levenshtein, Hamming ou Trigram par exemple.

Parmi les méthodes supervisées, le *boosting* [Freund et Schapire (1997)] réunit plusieurs avantages. Basé sur l'utilisation d'ensembles de prédicteurs ayant chacun un poids associé, il fournit naturellement une mesure de confiance dans ses prédictions.

Cette confiance peut d'abord être utilisée pour décider si une fusion entre lieux doit ou non être effectuée. Associée à un algorithme *clustering hiérarchique ascendant*, on contrôle ainsi la qualité des données. Au préalable, un algorithme de *k plus proches voisins* permet de réduire l'espace de recherche de lieux similaires.

La mesure de confiance fournie par le boosting peut aussi servir à proposer des exemples à étiqueter dans le cadre d'un *apprentissage actif* [Wang et al. (2009)]. Cela permet d'améliorer les résultats de l'algorithme d'apprentissage tout en minimisant l'effort requis par l'expert chargé d'étiqueter les exemples d'apprentissage.

3 Recherche d'Information Personnalisée

À ce niveau, on a donc à disposition un ensemble de lieux, chacun étant rattaché à une zone géographique et à un ensemble de catégories, organisées de manière hiérarchique.

L'étape suivante consiste à *interpréter les requêtes* utilisateurs : «resto brest» par exemple. Plusieurs problèmes se posent à ce niveau. Une *correction orthographique* doit être appliquée. Des suggestions peuvent être présentées pour les requêtes ambigües comme «boutique orange» ou «cafe de france».

Ensuite, étant donné un ensemble de lieux répondant à une requête utilisateur, un ordre dans la présentation des résultats doit être établi. On touche alors à la problématique importante du *learning-to-rank* [Cao et Liu (2007)]. De nombreux paramètres entrent dans la fonction permettant de prendre cette décision :

– adéquation entre le lieu et la requête
– qualité intrinsèque du lieu (présence d'informations plus ou moins complètes)
– qualité perçue du lieu (notations fournies par les utilisateurs)
– position géographique vis à vis de l'utilisateur (plus ou moins éloigné)

Enfin deux autres paramètres de personnalisation peuvent être considérés :
– la correspondance entre le lieu et le profil utilisateur (l'ensemble de ses goûts déclarés)
– la proximité avec les goûts de ses amis (facebook)

On touche ici aux sujets récents des *systèmes de recommandations* [Candillier et al. (2008)] et des *réseaux sociaux*, qui peuvent aussi être utilisés pour suggérer de nouveaux lieux aux utilisateurs sans qu'ils aient besoin de préciser de critère de recherche.

4 Perspectives

Toutes ces problématiques présentées ont été abordées par l'équipe Nomao, qui propose aujourd'hui un moteur mature et reconnu par ses utilisateurs. Mais les études peuvent bien sûr être davantage approfondies. Plusieurs sont en cours et pourront donner lieu à des soumissions d'articles scientifiques. Une réflexion est également en cours sur l'organisation de *challenges*, qui permettraient de mettre à disposition de la communauté scientifique des jeux de données réels à analyser sur des sujets porteurs.

Références

Candillier, L., K. Jack, F. Fessant, et F. Meyer (2008). *State-of-the-Art Recommender Systems*, Chapter 1, pp. 1–22. Collaborative and Social Information Retrieval and Access : Techniques for Improved User Modeling. IGI Global.

Cao, Z. et T.-Y. Liu (2007). Learning to rank : From pairwise approach to listwise approach. In *Proceedings of the 24th International Conference on Machine Learning*, pp. 129–136.

Freund, Y. et R. E. Schapire (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences 55*(1), 119–139.

Sarawagi, S. et A. Bhamidipaty (2002). Interactive deduplication using active learning. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 269–278.

Wang, Z., Y. Song, et C. Zha (2009). Efficient active learning with boosting. In *Proceedings of the 9th SIAM International Conference on Data Mining*, pp. 1232–1243.

Summary

Nomao is a search engine of places that ranks results according to what you like and your social network. Its development raises many scientific issues : extraction and structuralization of local content, query understanding and information retrieval, results ranking, personalisations and recommendations.