Titre de la thèse

Extraction de trajectoires dans les flux de données de mobilité : indexation dynamique et distribuée

Directeur de thèse : Stéphane Gançarski

Co-encadrant : Hubert Naacke

Mots-clés :

1. Systèmes de gestion de données à large échelle
2. Flux de données
3. Optimisation de requêtes spatiales parallèles et distribuées
4. Structures de données et indexation
5. Web sémantique et bases de connaissances
6. Expérimentation Apache Spark

Résumé

Dans le contexte de la gestion de grands flux de données géo-localisées issus des réseaux sociaux, le besoin est fort d'obtenir des informations sémantiquement plus riches sur les actions et déplacements des utilisateurs. Le croisement de grands flux de données avec des grandes bases de connaissances pose plusieurs défis lorsqu'on doit faire face à la dynamicité et au biais extrêmement important des données. Cette thèse a pour objectif d'extraire des trajectoires dans des flux de données de mobilité, puis de concevoir une solution capable d'analyser des trajectoires à large échelle. Un modèle de trajectoire sera défini, des structures de données et des méthodes d'indexation seront conçues pour manipuler efficacement ces trajectoires. Les algorithmes proposés seront implantés et validés expérimentalement en tirant profit des capacités de calcul parallèle de la plateforme Apache Spark. Les résultats attendus permettront, plus généralement, une meilleure gestion des données spatiales dynamiques et biaisées. Cette thèse pourrait avoir un impact socio-économique en améliorant les usages des réseaux sociaux.

Contexte et objectifs

Aujourd'hui les réseaux sociaux en ligne ont été adoptés massivement par une large tranche de la population, et sont devenus un outil quotidien de communication. Du fait de la géolocalisation, les données générées par ce type d'application sont des faits qui associent le plus souvent une personne, un lieu, une date et un contenu. Afin de mieux comprendre les nouveaux usages que les utilisateurs inventent avec de tels outils, le besoin d'analyser les données des réseaux sociaux est crucial et sans précédent. On constate que les analyses souhaitées sont de plus en plus complexes et concernent des données qui arrivent continuellement en quantité de plus en plus grande. De plus, avec la maturité et l'accessibilité des grandes bases de connaissances sémantiques telle que Wikidata [6], il y une demande forte pour enrichir les données des réseaux sociaux avec des informations sémantiques qui offrent l'opportunité de produire des résultats d'analyse plus détaillés et surtout plus explicables.

Dans le cas de la recommandation touristique, il s'agit d'indiquer à des voyageurs quels sont les points d'intérêt correspondant à leurs goûts et localisé à proximité ou sur leur itinéraire prévu. Cette fonction est

habituellement réalisée par des algorithmes d'apprentissage [2][3]. La plupart des solutions récentes [4] se limitent à considérer une faible quantité de données couvrant une zone relativement petite (ville ou région) ; elles ne passent pas à l'échelle. Nous avons récemment proposé une solution [1] scalable pour préparer une grande masse de données en croisant les données géo-localisées de photographies annotées (le jeu de données Yahoo [5]que nous utilisons comporte 100 millions de métadonnées couvrant tout le globe) avec la base Geonames [8] qui permet de distinguer les catégories des points d'intérêt. Notre contribution principale est une nouvelle méthode pour calculer des jointures spatiales plus rapidement en optimisant GeoSpark [7] afin de dédier plus de ressources aux points d'intérêt les plus populaires. Ce travail est en collaboration avec l'université de Thiès au Sénégal.

Pour aller plus loin, cette thèse a pour objectif de disposer d'informations précises sur les déplacements des individus et sur les points d'intérêt pertinents vis-à-vis de ces déplacements, et pouvoir ainsi mieux expliquer et anticiper les déplacements. Cela soulève des défis non résolus pour stocker et interroger efficacement des flux de données à large échelle, par exemple en comparant des trajectoires de différents utilisateurs afin de créer des profils génériques.

Les résultats attendus ont des applications dans des domaines importants et variés tels que la recommandation de point d'intérêt ou l'aide à la décision pour la mise en place et l'optimisation de réseaux de transports.

Problématiques de la thèse

Cette thèse aborde les problèmes posés par l'extraction et l'analyse de trajectoires dans les flux de données mobilité issues des réseaux sociaux.

L'un des problèmes est que ces données sont dynamiques : de nombreuses trajectoires sont insérées dans la base continuellement. De nouvelles trajectoires apparaissent et d'autres se complètent. C'est pourquoi la solution proposée doit être incrémentale.

Un autre problème est que les données sont fortement biaisées, certains lieux étant extrêmement plus visités que d'autres. Cela nécessite une allocation des ressources spécifiques pour stocker et interroger ces lieux.

Pistes

Il s'agira dans un premier temps de définir un modèle de trajectoire tenant compte de contraintes spatio-temporelles (par exemple une trajectoire dure moins d'une semaine et couvre une zone de moins de 10km) et d'informations sémantiques sur les utilisateurs (goût,...) et les lieux (catégories, ...).

Dans un deuxième temps, une structure de donnée et un algorithme seront proposés pour calculer efficacement ces trajectoires de manière incrémentale sur des données en constante évolution. Cela demandera notamment de concevoir une solution d'indexation dynamique et distribuée des données.

Dans un troisième temps, il s'agira de caractériser des types de requêtes d'analyse avec agrégation sur ces trajectoires (par exemple calcul de trajectoires fréquentes, calcul de similarité entre trajectoires). Après avoir démontré que ces requêtes d'analyse ne peuvent pas être exprimées entièrement en SQL, le langage d'interrogation sera étendu, de nouveaux opérateurs seront définis et les algorithmes d'évaluation associés seront conçus. Une attention particulière sera portée sur le traitement des données biaisées et leur impact sur les performances des requêtes parallèles : il s'agira de maximiser le degré de parallélisme effectif.

Les travaux de thèse conduiront à définir et implanter une architecture efficace pour l'extraction et l'analyse de trajectoires à large échelle. La plateforme expérimentale envisagée est Apache Spark qui est aujourd'hui la référence en termes de calcul parallèle sur cluster. Un des résultats attendus est d'étendre la plateforme Spark avec des nouveaux opérateurs de jointure par similarité qui soient robustes aux données biaisées.

Bibliographie

[1] Ibrahima Gueye, Hubert Naacke, and Stéphane Gançarski. Enriching geolocalized dataset with POIs descriptions at large scale. In Conference on Research in Computer Science and its Applications (CNRIA). Bambey, Senegal, 2020.

[2] Jean-Benoît Griesner, Talel Abdessalem, and Hubert Naacke. POI Recommendation: Towards Fused Matrix Factorization with Geographical and Temporal Influences. In 9th ACM Conference on Recommender Systems, RecSys. ACM, Vienne, Austria, 301–304, 2015.

[3] Jean-Benoît Griesner, Talel Abdessalem, Hubert Naacke, and Pierre Dosne. ALGeoSPF: Un modèle de factorisation basé sur du clustering géographique pour la recommandation de POI. In Extraction et Gestion de Connaissances (EGC 2018), France, 2018.

[4] K. H. Lim, J. Chan, S. Karunasekera, and C. Leckie. Personalized itinerary recommendation with queuing time awareness. In ACM SIGIR Conference on Research and Development in Information Retrieval, pages 325–334, 2017.

[5] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth,and L.-J. Li. YFCC100m: The new data in multimedia research. Commun. ACM,59(2):64–73, 2016.

[6] Vrandecic and M. Krotzsch. Wikidata: A free collaborative knowledgebase. Communications of the. ACM, 57(10):78–85, 2014.

[7] J. Yu, J. Wu, and M. Sarwat. Geospark: a cluster computing framework for process-ing large-scale spatial data. InSIGSPATIAL International Conference on Advances in Geographic Information Systems, pages 70:1–70:4, 2015.

[8] Geonames. The geonames dataset."http://www.geonames.org/export/".