MANIE Mohamed
P0908132

Application de mise en correspondance d'entités avec pré-sélection par Map Reduce

Année 2013-2014

I) Présentation du projet de recherche

Le web permet l'accès à un très grand nombre d'informations (documents HTML, données relationnelles, API, entités sémantiques,…). Dans le cadre du projet KOGAR[1], nous nous focalisons sur des entités sémantiques dans un contexte où différents sites peuvent disposer d'entités différentes qui décrivent les mêmes concepts. Pour relier les entités décrivant le même concept, différentes techniques d'« entity matching » [1] existent et s'avèrent reposer sur des algorithmes coûteux en temps et ressources de calculs. Dans le cadre du projet, nous voulons enrichir des collections d'entités, pour cela il est d'abord nécessaire de détecter quelles sont les entités équivalentes entre les différentes collections. Cependant, le nombre d'entités dans chaque collection est très important et les comparer deux à deux engendrerait des latences inacceptables pour les utilisateurs.

L'approche choisie dans ce projet repose sur le paradigme du Map Reduce [2] qui est une méthode de développement qui permet d'effectuer des calculs parallèles. Popularisé par Google pour l'implémentation de son index, l'intérêt de cette méthode est de manipuler de grandes quantités de données en les distribuant dans un cluster de machines pour être traitées. L'algorithme repose sur une première étape qui consiste à découper un problème en plusieurs sous problème, suivie d'une seconde étape permettant de faire remonter et d'agréger les résultats des nœuds fils vers les nœuds pères.

Des travaux existants dont *Load Balancing for MapReduce-based Entity Resolution* [3] montrent la nécessité de passer par une phase de « blocking » pour faire de l' « entity matching ». Cette phase consiste à découper les données en entrée en des blocs plus petits de tailles identiques. Ensuite, ces blocs sont attribués à des nœuds disponibles pour leur faire exécuter les alignements sur ce bloc, qui consiste à attribuer à chaque couple <clé, valeur> obtenu à l'aide de la fonction **emit(),** un ensemble de nouveaux couples list(clé, valeur).

L'objectif du stage consiste à définir un algorithme à la Map-Reduce intégrant en amont une phase de blocking pour permettre la découverte de correspondances sémantiques entre différentes entités (voir figure1).

---

[1] Projet KOGAR : http://liris.cnrs.fr/~fduchate/projets/aurora

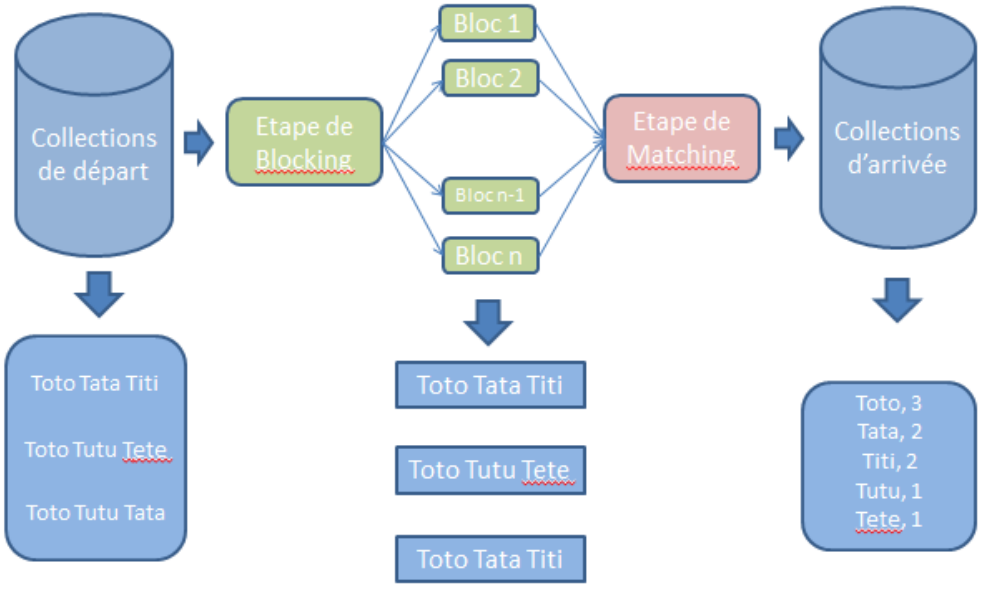

Figure1 : Aperçu du processus « entity matching »

II) Aspect technique du projet

Le prototype qui permettra de sélectionner les différentes entités à matcher devra être codé en JAVA afin de s'intégrer dans une application existante développée dans le cadre du projet KOGAR. Pour cela, il sera nécessaire de tenir compte du modèle de données existant. L'algorithme proposé devra être testé et validé.

III) Déroulement du projet

Pour mener à bien ce projet, je vais devoir effectuer les tâches suivantes :

- Rédiger un cahier des charges.
- Lire plusieurs articles sur « entity matching » et sur « Map reduce » afin de comprendre la mise en correspondance d'entités en utilisant un paradigme à la Map Reduce.
- Comprendre les classes et la logique utilisée dans l'application du projet KOGAR[1].
- Proposer un scenario pour la conception du prototype.
- Travailler ce scenario et un pseudo code en fonction des remarques des encadrants.
- Débuter l'implémentation du prototype.
- Tester le prototype sur différentes données.
- Comparer les résultats pour estimer l'efficacité de l'algorithme.

IV) Calendrier prévisionnel

La Figure2 présente le calendrier prévisionnel pour le déroulement de ce projet.

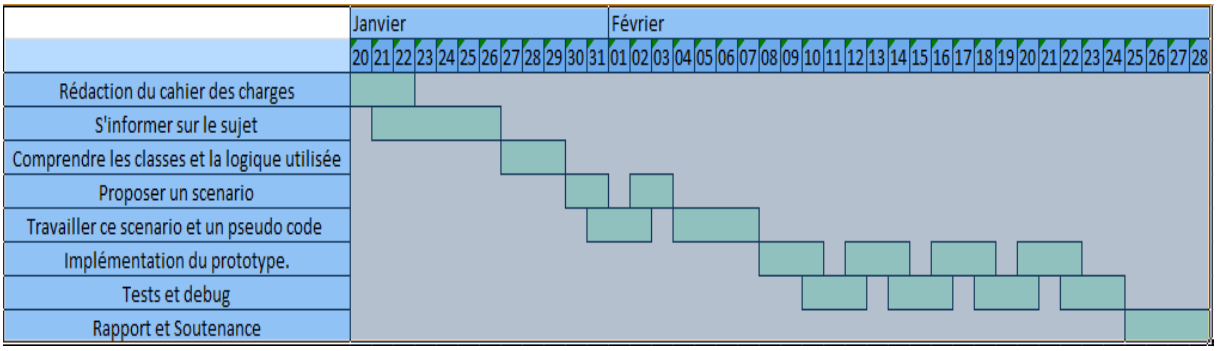

Figure2 : Diagramme de Gantt

V) Références

[1]: Tamer Elsayed, Jimmy Lin, and Douglas Oard: *Pairwise Document Similarity in Large Collections with MapReduce.* Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL 2008), Companion Volume, pages 265-268, June 2008, Columbus, Ohio.

[2]: Jeffrey Dean and Sanjay Ghemawat. MapReduce: *simplified data processing on large clusters.* In *OSDI'04: Proceedings of the 6th conference on Symposium on Opearting Systems Design & Implementation*, Berkeley, CA, USA, pages 137-149, 2004. USENIX Association.

[3]: Lars Kolb, Andreas Thor, Erhard Rahm*: Load Balancing for MapReduce-based Entity Resolution*. ICDE 2012: 618-629.