PlantRT : a Distributed Recommendation Tool for Citizen Science

Maximilien Servajean, Esther Pacitti, Miguel Liroz-Gistau, Alexis Joly, Julien Champ

PlantRT : a Distributed Recommendation Tool for Citizen Science

PlantRT : Outil de recommandation distribué pour les sciences citoyennes

Maximilien Servajean
INRIA & LIRMM
University of Montpellier
France
servajean@lirmm.fr

Esther Pacitti
INRIA & LIRMM
University of Montpellier
France
pacitti@lirmm.fr

Miguel Liroz-Gistau
INRIA & LIRMM
Miguel.Liroz_Gistau@inria.fr

Alexis Joly
INRIA & LIRMM
alexis.joly@inria.fr

Julien Champ
INRIA & LIRMM
julien.champ@lirmm.fr

ABSTRACT

Les utilisateurs du Web 2.0 sont de gros producteurs de données diverses qu'ils stockent dans une grande variété de systèmes. Dans ce travail, nous nous concentrons sur le cas particulier des botanistes. En effet, établir une connaissance précise de l'identité, de la distribution géographique et de l'évolution des espèces vivantes est essentiel pour la pérennité de cette biodiversité, tout autant que pour l'espèce humaine. L'émergence des sciences citoyennes et des réseaux sociaux sont des outils supplémentaires favorisant la création de grandes communautés d'observateurs de la nature, qui ont commencé à produire d'énormes collections de données multimédias. Cependant, la complexité inhérente à la réalisation de ces collections provoque une certaine méfiance des utilisateurs, ces dernier ne souhaitant pas stocker leurs données sur un serveur central. Dans ce travail, nous avons réalisé un prototype multi-sites, où chaque site, peut représenter 1 à n utilisateurs permettant la recherche et la recommandation d'observations de plantes diversifiées à grande échelle.

Categories and Subject Descriptors

H.4 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Search process

Keywords

Multi-sites, top-k, search and recommendation

1. INTRODUCTION

Web 2.0 users are massive producers of diverse data (*e.g.* photos, videos, scientific data). Additionally, while users are often willing to share their data with each other in a community of interest, they do not want to lose control over them using a central site. However, the distribution of the users' data in many different devices (*e.g.* own computer, servers) makes data sharing especially difficult.

In this work, we focus on the particular case of botanists. Building an accurate knowledge of the identity, the geographic distribution and the evolution of living species is essential for a sustainable development of humanity as well as for biodiversity conservation. The emergence of citizen sciences and social networking tools has fostered the creation of large and structured communities of nature observers (*e.g.* e-bird, xeno-canto, Tela Botanica, Pl@ntNet [3]) who started to produce outstanding collections of multimedia records. Scaling up such collaborative approaches to real-world ecological surveillance systems involving millions of contributors is however still challenging. In this context, users, or botanists, make observations of plants. An observation is composed of the plant's picture and its associated metadata, namely, the plant family, genus and species, the observation's geographic position (*i.e.* GPS) and time, and a description. The effort required to build a high-quality collection of observations is the reason why some botanists want to keep their data on their own computers or small servers. However, they still want to share observations, though in a controlled manner, so that they can search and be recommended from the whole community's observations.

Problem Definition: the goal of our work is to propose a large scale distributed platform that enables searching and recommending relevant and diversified plants observations taking into account both items' content and users' profile.

2. ARCHITECTURE'S OVERVIEW

In our distributed search and recommendation approach, users are represented by virtual nodes. Each site is composed of 1 to p virtual nodes and a virtual node can be composed of 1 to m users with a similar profile. Users can select the site to which they are connected, while they are

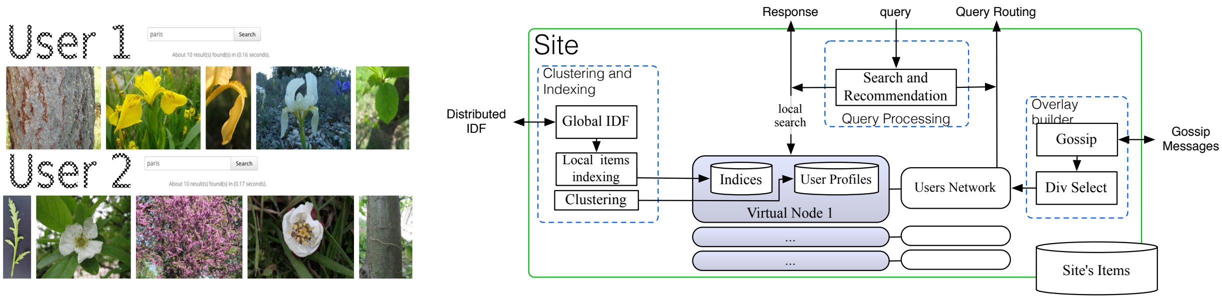

Figure 1: PlantRT's search and recommendation example and architecture.

clustered to virtual nodes depending on their profile. Each virtual node is associated to an index containing all items stored in the site. Items are indexed with respect to the profiles of the users associated to the current virtual node, so within a site all indexes will point to the same set of elements but will rank them in a different order. Virtual nodes are connected between them through a *User Network* overlay.

Whenever a user u submits a query q, the system sends it to this subset of virtual nodes (*i.e.* User Network), that will return their relevant results to u and will also recursively forward the query to the virtual nodes in their respective *User Network* until the *TTL* is reached. To build *User Network*, we use a two step approach. First, based on *random gossiping*, each site s is aware of other virtual nodes available on the network. Second, by means of a *diversified clustering* algorithm, u's virtual node chooses among these virtual nodes the best ones to answer u's queries and keep them in *User Network*.

Thus, PlantRT uses three components: (1) PlantRT Clustering and Indexing which is in charge of managing the virtual nodes and data indexing, (2) PlantRT User Network which is in charge of establishing the overlay between virtual nodes among sites and (3) PlantRT Query Processing, which is in charge of propagating queries submitted by users through a *User Network* overlay, to a subset of relevant virtual nodes and processing them. PlantRT's architecture is presented in Figure 1.

2.1 PlantRT Clustering and Indexing

The first component, deals with three tasks: a) maintaining the virtual nodes with similar users, b) maintaining the indices employed to retrieve items from keywords and geo-position in each virtual node, and c) generating the users profile from the locally shared items.

Based on [1], similar users are clustered together, each cluster corresponding to a virtual node. This is done by executing periodically *k-means* clustering algorithm. An index of all items stored in the site is built taking into account the profiles of the users in the cluster. The maximum number of virtual nodes is system defined and depends on the storage and memory capacity of the site [1].

Items (*i.e.* observations) are associated to both their GPS position and to a keywords vector [5] where each of them is associated to a *tf* $\times$ *idf* (*i.e.* term frequency $\times$ inverse document frequency) score representing its importance in the item with respect to the whole corpus I.

However, unlike centralized solutions, since the global corpus is not available at each site, PlantRT uses a gossip-based protocol that progressively produces the score of each keyword in the set of items shared in the site s with respect to the distributed global corpus of items I.

The intuition behind our distributed *tf* $\times$ *idf* protocol is that statistics about the global corpus can be estimated using average computing which can be quickly computed using gossip protocols [4].

Then, a profile is only computed as the average of its shared observations' *tf* $\times$ *idf* vector. This information is used during index generation. Notice that this index is later used during query processing to *efficiently* recommend relevant items with respect to a query at each involved virtual node.

2.2 PlantRT User Network

The second component aims at establishing an overlay between virtual nodes. Based on random gossiping [2], each site s maintains a set of random view entries corresponding to the virtual nodes profile s is aware of. Periodically, sites gossip, and exchange a random subset of virtual nodes views entries. After the random gossip merging phase, a diversified clustering algorithm is triggered at each virtual node. In fact, taking into account the previous gossip exchange, the algorithm selects the most relevant nodes – using a similarity measure – from the random view considering the relevant nodes previously selected in the *User Network* of each virtual node. Indeed, the *User Network* should diversified to increase coverage and therefore the probability to answer any query [citation globe].

2.3 PlantRT Query Processing

Finally, the third component deals with the execution of queries. Whenever u submits a query q, the query is redirected to all virtual nodes in the participating nodes' *User Network* recursively, until a predefined upper threshold, *TTL* (*i.e. Time-To-Live*). Whenever a node v receives a query, it computes its *top-k* most relevant and diversified items, taking into account both items' content and users' profile [6], among the locally indexed elements with respect to the query using a specific similarity measure (*e.g.* jaccard). Then, v returns its set of recommended items to u. An item recommended by a user v_i is defined by its identifier, its score, the site's identifier and v_i's profile. Once u receives the set of recommended items from all users $v_1, ..., v_n$ that received the query q, it ranks all received recommendations based on their score and on the similarity of v_i with respect to u.

3. PLANTRT DEMONSTRATION

Figure 1 presents the results of a search executed on our prototype. It shows a use case where two users are searching plants around Paris. However, since they are not interested in the same kind of plants both results lists are different. Also, the results are diversified in the sense that they only contain plants from different families.

Our prototype can be deployed on several nodes. In our experiments we have simulated up to $6,000$ nodes, reaching a recall (*i.e.* the proportion of results answering a query retrieved) of $99,9\%$ [citation globe].

4. CONCLUSION

This work presents the implementation of a diversified and distributed recommendation tool for citizen science, and more precisely, for botanists. Our platform integrates a complete set of features (*e.g.* subscribe, share, search, recommendation) and the still evolving source code is available online [1]

1. `http://www2.lirmm.fr/~servajean/prototypes/plant-sharing/plant-rt.html`.

5. REFERENCES

[1] S. Amer-Yahia and M. Benedikt. Efficient network aware search in collaborative tagging sites. *VLDB Endowment '08*, 1(1):710–721, 2008.

[2] M. Jelasity and O. Babaoglu. T-man: Gossip-based overlay topology management. In *ESOA*, volume 3910 of *Lecture Notes in Computer Science*, pages 1–15, Berlin, Heidelberg, 2005.

[3] A. Joly, H. Goëau, P. Bonnet, B. Vera, J. Barbe, S. Souheil, Y. Itheri, J. Carré, E. Mouysset, J.F. Molino, N. Boujemaa, and D. Barthélémy. Interactive plant identification based on social image data. In *Ecological Informatics*, 2013.

[4] W. Kowalczyk, M. Jelasity, and AE. Eiben. Towards data mining in large and fully distributed peer-to-peer overlay networks. In *BNAIC*, pages 203–210, 2003.

[5] G. Salton. *Automatic Information Organization and Retrieval*. McGraw Hill Text, 1968.

[6] M. Servajean, E. Pacitti, S. Amer-Yahia, and P. Neveu. Profile diversity in search and recommendation. In *WWW Companion*, pages 973–980, 2013.