A PERSONALIZED AND DIVERSIFIED MODEL FOR DECENTRALIZED SCIENTIFIC SEARCH AND RECOMMENDATION

Maximilien Servajean (Doctorant)
INRIA & LIRMM
Université de Montpellier
servajean@lirmm.fr

Esther Pacitti
INRIA & LIRMM
Université de Montpellier
pacitti@lirmm.fr

Sihem Amer Yahia
CNRS, LIG
sihem.amer-yahia@imag.fr

Pascal Neveu
INRA & Supagro
pn@supagro.fr

LABEX

NUMEV Solutions Numériques Matérielles et Modélisation pour l'Environnement et le Vivant

Positionnement NUMEV : Axe Données

Keywords : recommendation, top-k, diversity, decentralized systems, plant phenotyping

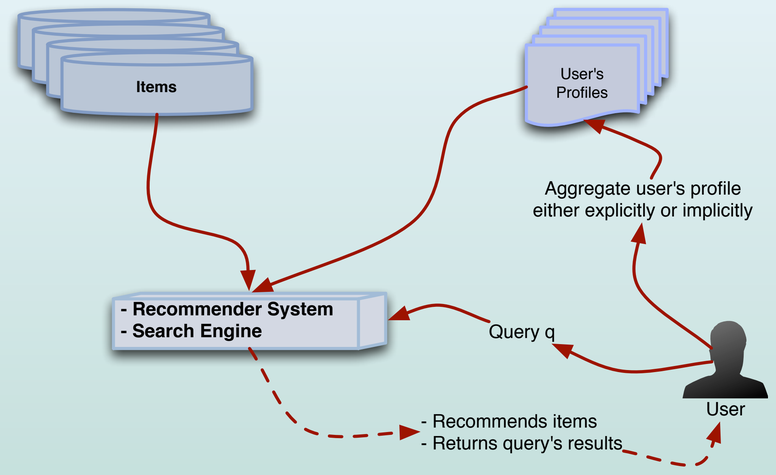
Abstract: Increasingly, scientific progress depends on exploring existing results and correlating innovations in different research communities to derive novel solutions to challenging problems. In fact, in the domain of plant phenotyping there has recently been increasing interests in finding indirect associations between concepts in different published papers coming from different research communities to derive new methods. We investigate profile diversity, a novel idea in searching scientific documents. Combining keyword relevance with popularity in a scoring function has been the subject of different forms of social relevance. Content diversity has been thoroughly studied in search and advertising, database queries, and recommendations. In our approach we investigate profile diversity to address the problem of returning highly popular but too-focused documents. To handle this problem we adapt Fagin's threshold-based algorithm to return the most relevant and most popular documents that satisfy content and profile diversities constraints. We also exploit diversification in a decentralized scenario, which is a typical set up in the plant phenotyping community.

APPROACH:

We use recommendation and information retrieval techniques

The results returned to a user depends on:

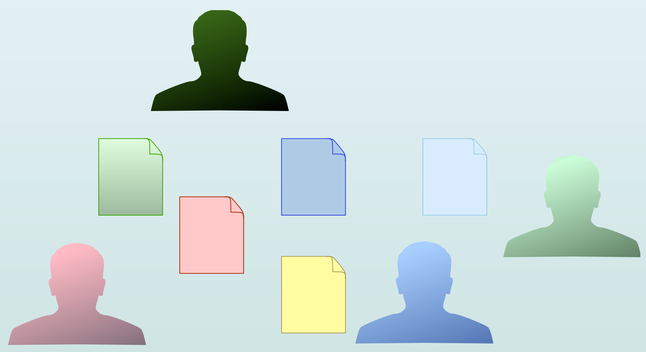
- the whole set of users
- the query's initiator
- the query
- the set of items



DIVRSCI: A DIVERSIFIED AND PERSONALIZED MODEL

DivRSci returns relevant and diversified documents shared by relevant and diversified users:

$$score_{DivRSci}(d, u, q) = rel(d, q).div_c(d|\{d_1, ..., d_{i-1}\}).div_p(u_d|\{u_{d_1}, ..., u_{d_{i-1}}\})$$



DivRsci uses a new Profile Diversification score:

$$div_p(u_d|\{u_{d_1}, ..., u_{d_{i-1}}\}) = \frac{1}{N} \cdot \sum_{v_n \in u_{d_i}} \left[rel_{trust}(v, u, q). \prod_{v_m \in \{u_{d_1}, ..., u_{d_{i-1}}\}} (1 - red_p(v_m|v_n) \right]$$

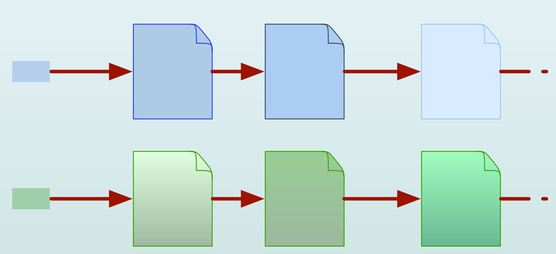
Profile Diversification enables to return documents shared by trustworthy and diversified users with respect to the query's initiator and to the query itself

DIVRSCI: A DIVERSIFIED THRESHOLD ALGORITHM

DivRsci uses a threshold algorithm which iterates over a set of inverted index. The algorithm stops when a threshold condition is satisfied

We propose a new Threshold to reduce the number of iterations during query processing:

$$\delta' = f(s_1, s_2, ..., s_n).f_{div_c}(d_i, \{s_1, s_2, ..., s_n\}).f_{div_p}(d_i, \{s_1, s_2, ..., s_n\})$$



Inverted Index

Results

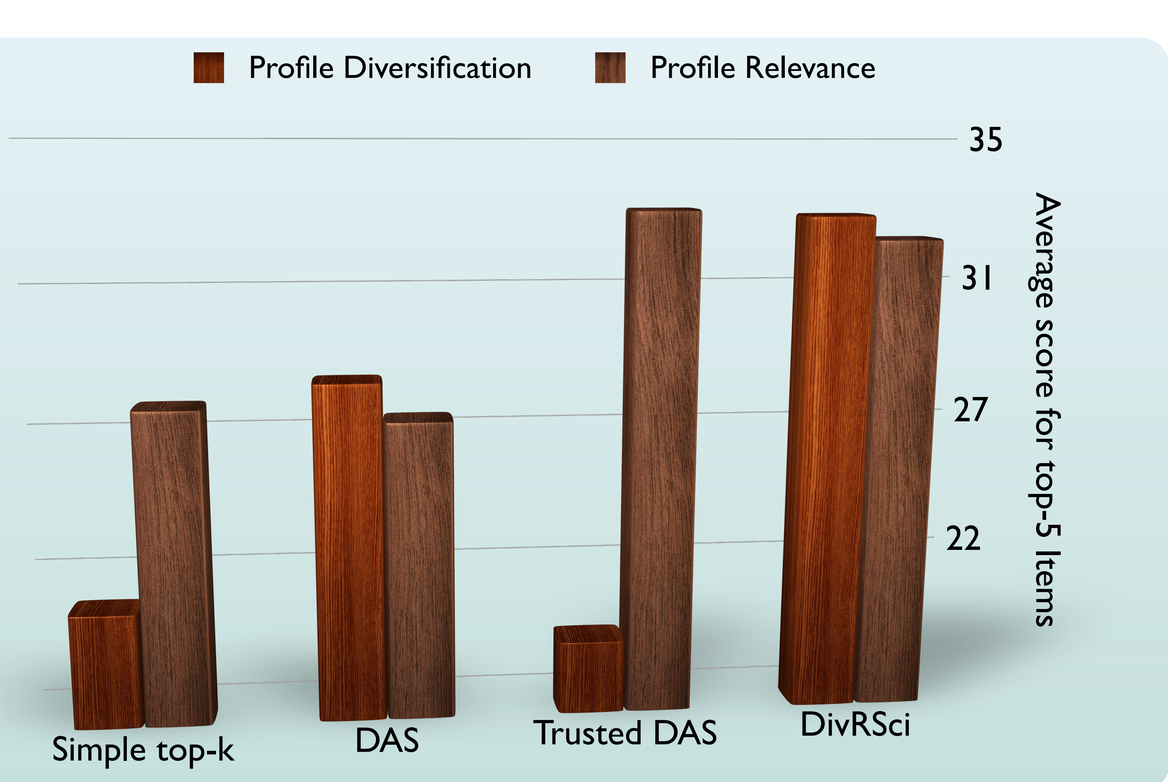
Simple top-k only takes in account similarity with the query

DAS takes in account similarity with the query and diversification

Trusted DAS takes in account similarity with the query, diversification and personalization

DivRSci is the complete model

DivRSci enables to have a better compromise between Profile diversity and profile relevance than other scoring functions. In other words, *DivRSci* enables to retrieve documents from both various and relevant communities or disciplines.



[1] M. Servajean, E. Pacitti, S. Amer-Yahia, and P. Neveu, "Profile Diversity in Search and Recommendation" SRS'13, 2013.