Paper Recommendation System: A Global and Soft Approach

Siwipa Pruitikanee, Lisa Di Jorio, Anne Laurent, Michel Sala

Paper Recommendation System: A Global and Soft Approach

Siwipa Pruitikanee, Lisa Di Jorio, Anne Laurent, Michel Sala
LIRMM - Univ. Montpellier 2 CNRS
Montpellier, France
Email: laurent@lirmm.fr

***Abstract*—Paper recommendation to researchers has been extensively studied in the last years, and many methods have been investigated for this purpose. In this paper, we propose a novel approach embedding the whole process for selecting papers of interest given some keywords. Our approach is based on a workflow integrating fuzzy clustering of the papers, the computation of a representative summary paper per cluster using OWA operators, and ranking, in order to answer user queries adequately. The originality of our method relies in the introduction of fuzziness for more flexibility in the approach. The use of representative papers allows us to summarize sets of papers into a single representative one, thus simplifying the users interactions with the huge number of papers from the literature.**

***Keywords*-Paper Recommandation; Soft Computing; Fuzzy Clustering; OWA**

I. Introduction

As the scientific literature is growing dramatically, selecting and reading papers has become a hard task, especially in the case of literature review. Digital libraries provide tools to help the user navigate through the resources and query the datasets. We discern many reasons for choosing and reading a paper; among them are the need to be aware of every new potential discovery in very specific domains, or the paper selection in a literature review process, as for example when writing an academic paper. In this context, recommending papers meeting some criteria such as the conference or author ranking is of great importance in order to avoid the time consuming step of reading many papers that are not so relevant to the subject.

Most of Digital libraries propose navigation tools, most of them based on multicriteria filtering and/or collaborative filtering. For this purpose, paper recommendation systems have been extensively studied in the last years. Some tools have been created to group very similar papers using clustering methods, to provide organised information to the user. However, none of these tools is able to point out representative papers. Thus, the reader has no idea of the main methods described in these groups of papers and of the most representative of these methods.

In this paper, we propose a novel approach embedding the whole process. Our approach is based on a workflow composed of four steps.

The first step consists in selecting papers that are related to the user query. In this step, all papers containing at least one keyword among those included in the user query are selected.

The second step consists in grouping papers based on their similarity. For this purpose, we consider that papers are similar if they deal with the same topics. As we consider that it is not relevant to split objects in a crisp manner, we consider here fuzzy clustering.

The third step consists in computing representative papers, allowing us to resume sets of papers into one, thus simplifying the user interactions with the huge number of papers from the literature. We propose this representative paper to get enriched by a small number of other papers from the group, in order to cover all the topics of the user query.

The fourth step consists in ranking the representative papers so as to present the papers to the user in decreasing order of interest. In this step, we consider classical methods, such as PageRank.

The originality of our approach is twofold: in one hand we consider the whole workflow and on the other hand we introduce fuzziness in order to soften the approach.

The paper is organized as follows. Section II presents the existing work related to our approach. Sections III and IV introduce the running example and the formal framework we rely on in the proposition Section V details our proposition. Finally, Section VI draws some conclusions on our work and proposes research directions.

II. Related Work

Paper recommendation systems lie at the intersection of different fields of data analysis: recommendation systems, text ranking and scientometry. In this section, we discuss of the main advances in each domain, and of the main drawbacks.

A. Paper Ranking Methods

Large efforts have been provided regarding the ranking of papers. Papers can be evaluated and compared using different criteria: the authors reputation, the date of publication, the conference or journal ranking and the number of times it has been cited, are among the most often used information.

Citation count is one of the most used information. For example, the Thomson Scientific ISI impact factor (ISI IF) [1] is based on citation counts. It combines citation counts with a moving window to favor the most recent papers, and also include the impact of some journals in the calculation. However, methods based on the citation count suffer from the fact that paper impact is not taken into account. Thus, recent works have proposed a modified version of the ISI IF to integrate the "popularity factor", which is defined by the citation analysis of publication venues and the PageRank algorithm. [2] modified the PageRank algorithm in order to apply it on academic papers. As in PageRank, the quality of a paper is based on the number of papers pointing to it, and its quality decrease if there are too much outgoing links (citations) from it. However, this approach suffers from the following drawback: some good papers (especially survey papers) need a lot of citation in order to contextualize their work. Moreover, as for PageRank, the algorithm has some difficulties to take very recent papers into account, no matter their quality.

B. Recommender Systems

Recommender systems are an important research field since the 90's, mainly because of their generic and industrial application. Roughly speaking, a recommender system takes some user interest or profile as an input, and searches among massive database information for items that the user has not seen and which he may be interested in. Recommender systems differ from classical data mining as it has to deal with specific user profile and result ranking.

There are two main paradigms in Recommender Systems: Collaborative Filtering (CF) and Content-Based Filtering (CBF). Note that some works propose "hybrid" approaches by mixing the two paradigms. The interested reader may refer to [3] for a detailed survey about these different approaches.

In Collaborative Filtering ([4], [5]), the systems propose items to a user, by considering the items that similar users liked in the past. Thus, CF systems rely on rating and profiling. Such systems are quite mature and currently used in e-commerce websites, for example Amazon[1] or Ebay[2]. Among the weakness of such systems are the "cold start problem": when starting or adding new items, the system needs some elements to be initialised before being able to predict. When a new user is added, the system needs to profile him in order to make efficient recommendation. Finally, there is a sparsity problem, as there are only a small set of rates compared to the set of recommendation that has to be predicted.

In Content Based Filtering ([6], [7]), items that a user already pointed out as being of interest are used to recommend new items. Thus, the process can be seen as a classification task, where the training set is the user preferences. As it has been widely used in text-based context (internet, news,...) CBF systems mainly use information retrieval and information filtering methods. However, such systems can be limited by the problem of content analysis, because of the format of input items; while research reached a mature point concerning text-based documents, feature extraction from stream or video based document is much harder. Also, CBF systems are limited to what the user feeds them: they will never recommend items from another domain than those already rated by the user.

More recently, recommender systems have been extended to the paper recommendation context.

[8] proposes a hybrid approach that mixes CF and CBF. The authors detail a set of tools ranging from the simple CF system using k-nn algorithm and enriching data by adding cited papers to CBF using TF-Idf measure. Here, hybridation occurs by merging the results of CF and CBF. The author concludes that the hybrid system performs better than only CF or only CBF. [9] also proposes a hybrid approach based on graphs. It allows both for users and items integration in the system. The authors are then able to use classical graph search for extracting and recommending useful information. As [8], the authors show that hybrid approaches outperform CBF methods.

[10] proposes a system based on a new random walk process and the citation graph, called ItemRank. It is based on PageRank through its propagation and attenuation properties. In [11], a CF approach is done by clustering a subspace of papers. In this paper, the main goal is to apply the system to researchers working in the same laboratory. The originality of the method is the clustering algorithm that efficiently traverses the search space by subspace intersection. [12] describes a ranking-oriented CF system which extracts users's access logs as the training set. The system overcomes the cold start problem, however, weblogs stay noisy and not reliable data [13].

[14] uses different informations such as the title, the abstract, the sentences around a citation in order to build a citation recommender system. The main novelty is the user query form: it does not have to be a bibliography, it can also only be a document or some specific sentences.

In [15], a user can give as an input an entire document. The process then uses every contextual information such as the citation analysis, authors, sources, implicit and explicit ratings. Moreover, the authors propose to use the Distance Similarity Index (DSI) and the In-text Impact Factor (ItIF). The authors build a system combining all user-given information parameters (for example an h-index range for author reputation) and provide a graphical user interface.

[16] only use the citation graph in order to output a small-

[1] www.amazon.com

[2] www.ebay.com

sized set of relevant papers. They define measures working at two granularity levels: the Local Relation Strength measures the dependency between cited and citing papers, and then the Global Relation Strength captures the relevance between two papers in the whole citation graph. The Local Relation Strength relies on weighted parameters such as the number of times a paper is cited, and the number of times two papers are cited together, or the age of a publication. Then, the Global Relation Strength combines the Kratz measure [17] with the dependency in a citation link. [18] uses the user's recent research interests in order to recommend new papers. The work focuses more on the user profile: the author discriminates between junior researchers and senior researchers. The authors hypothesis is that contextual information about the user can provide evidence for recommendation. In this context, the information is provided by the user historical search. Then, the paper selection is driven by the prebuilt profile.

C. OWA

When aggregating information, many operators are available [19], as weighted average. The idea here is to combine N values into a single result. [20], [21] propose the OWA operator, defined as below.

Definition 1: A vector $v = (v_1, \dots, v_N)$ is a weighting vector of dimension N if $v_i \in [0,1]$ and $\sum_{i=1}^{N} 1$.

Definition 2: A mapping AM: $\mathbb{R}^N \rightarrow \mathbb{R}$ is an arithmetic mean of dimension N if $AM(a_1, \dots, a_N) = (1/N) \sum_{i=1}^{N} a_i$.

Definition 3: Let p be a weighting vector of dimension N. A mapping WM: $\mathbb{R}^N \rightarrow \mathbb{R}$ is a weighted mean of dimension N if $WM(a_1, \dots, a_N) = \sum_{i=1}^{N} p_i a_i$.

Definition 4: Let w be a weighting vector of dimension N. A mapping OWA: $\mathbb{R}^N \rightarrow \mathbb{R}$ in an ordered weighting average of dimension N if $OWA\ (a_1, \dots, a_N) = \sum_{i=1}^{N} w_i a_{\sigma(i)}$, where σ is a permutation such that $\forall i \in [1, N-1] a_{\sigma(i)} \geq a_{\sigma(i+1)}$.

D. Fuzzy Clustering

Clustering consists in grouping together observations sharing the same characteristics, but without the help of predefined classes. Clustering method appeared in the 70s, and if some specific context still need to be explored, there exist several mature methods to compute this result, such as hierarchical clustering, K-means, C-means,... Some methods consider that clusters can overlap. These last solutions are known as *fuzzy clustering*[22]. Every object then belongs to every cluster with a membership degree ranging from 0 to 1. (Fuzzy) Clustering is based on a distance measure which is used for describing to which extent two objects are similar.

Author Id	Author Name	H-Index
a1	John Martin	88
a2	John Smith	78
a3	Jack Jibb	17
a4	Mark Clark	7
a5	Luis Martinez	5
a6	Lora Davis	57
a7	Pen Green	25
a8	Frank Lee	0
a9	Home Sweet	8

Table I
EXAMPLE DATASET - AUTHORS

Fuzzy C-means is one of the most often used method. Let us consider n objects $x_1, \dots, x_n$ described over d attributes. The objective is to group these objects into k clusters, each cluster c_i ($i = 1, \dots, k$) being represented by its center v_i. Let $u_{i,j}$ be the degree of membership of the object x_i in the cluster c_j.

Let $|| * ||$ be any norm expressing the similarity.

$u_{i,j}$ is computed as:

$$u_{i,j} = \frac{1}{\sum_{k=1}^{c} \left(\frac{||x_i - v_j||}{||x_i - v_k||} \right)^{\frac{2}{m-1}}}$$

The algorithm relies on a iterative process that computes, for every object, the membership degree to every cluster and recomputing the center of the clusters. The degree of fuzziness of the process, impacting the overlapping rate of the clusters, is tuned using the m parameter.

III. RUNNING EXAMPLE

We consider the example detailed in tables I, III and **??**

In this example, we consider several papers that have been published on topics identified by keywords. These keywords can belong to more than one paper. These papers have been written by authors and cite some other ones.

The abstracts and titles allow us to identify keywords. For instance, let us consider paper $p1$ published in 1996, it is related to data mining, with the following abstract:

The mining of large databases is a very hot topic in database systems and machine learning. Companies have used some data mining techniques for understanding customer behavior on their data warehouse. This article provides a survey on the data mining techniques, classification and comparing of data mining techniques.

A first step is to transform citation table to a binary matrix.

IV. FORMAL FRAMEWORK

In this section, we present the seminal definitions for describing the data we are dealing with.

Let:

- $D = \{p_1, p_2, ..., p_m\}$ be a set of research papers
- $K = \{k_1, k_2, ..., k_n\}$ be a set of keywords

Papers	Paper Id	Authors	Year	Number of Citations
	p1	J. Martin, J. Smith	1996	3
	p2	J. Martin, J. Smith	2003	1
	p3	J. Martin, J. Jibb	2005	0
	p4	M. Clark, L. Martinez	2006	0
	p5	L. Davis, P. Green	1997	2
	p6	L. Davis	2002	0
	p7	F. Lee, H. Sweet	2004	0

Table III
EXAMPLE DATASET - PAPERS AND AUTHORS

Paper Topics	Paper Id	Title	Conf/Jal	Year
	p1	A survey of Data mining techniques	s1	1996
	p2	Data streams mining with a classifiers	s2	2003
	p3	Summarization k representative rules of frequent-pattern	s2	2005
	p4	Data mining in money laundering crimes	s2	2006
	p5	Selection of relevant features and examples in machine learning	s3	1997
	p6	Machine learning for automatic text classification	s4	2002
	p7	Using Data Mining to Develop The Expert System	s5	2004

Table IV
EXAMPLE DATASET - PAPERS

Conf/Jal Ranking	Name	Ranking
	s1	A
	s2	B
	s3	A
	s4	B
	s5	B

Table II
CONF/JAL RANKING

Paper Id	Cite to
p2	p1
p3	p1
p4	p1
p6	p5
p7	p5

Table V
EXAMPLE DATASET - CITATIONS

Paper Id	p1	p2	p3	p4	p5	p6	p7
p1	0	0	0	0	0	0	0
p2	1	0	0	0	0	0	0
p3	1	0	0	0	0	0	0
p4	1	1	0	0	0	0	0
p5	0	0	0	0	0	0	0
p6	0	0	0	0	1	0	0
p7	0	0	0	0	1	0	0

Table VI
EXAMPLE DATASET - CITATION MATRIX

- $A = \{a_1, a_2, ..., a_q\}$ be a set of distinct authors

These sets are mapped using the following functions:

- $W : D \to \mathcal{P}(A)$, where $W(p)$ returns the set of authors of paper $p \in D$
- $T : D \to \mathcal{P}(K)$, where $T(p)$ returns the set of keywords embedded in the title of paper p
- $Ab : D \to \mathcal{P}(K)$, where $Ab(p)$ returns the set of keywords embedded in the abstract of paper p
- $C : D \to \mathcal{P}(D)$, where $C(p)$ returns the set of papers cited by paper p

V. PROPOSITION

Our proposition relies on a four-step process, starting from papers from several sources (e.g, Web of Science, DBLP, local databases) and arriving to representative papers ranked regarding their interestingness, as shown in Figure **??**.

The data pre-process has been done by collecting publication or academic paper data from multi-sources into one data structure. Our structure focuses on common attributes, being composed of title, authors, published date, source (e.g., journal) and citations or reference list.

A. Step 1: Selecting Papers

The process starts with publication selection and is based on keywords provided by the user in her/his query. This step returns the papers that match at least one of these given keywords. We thus obtain the preliminary related publications dataset. For instance, let us consider a user choosing the following two keywords: data mining and machine learning. Both of them are separated into 4 given words: data, mining, machine and learning, and use them for finding the publications from database, assume the result is

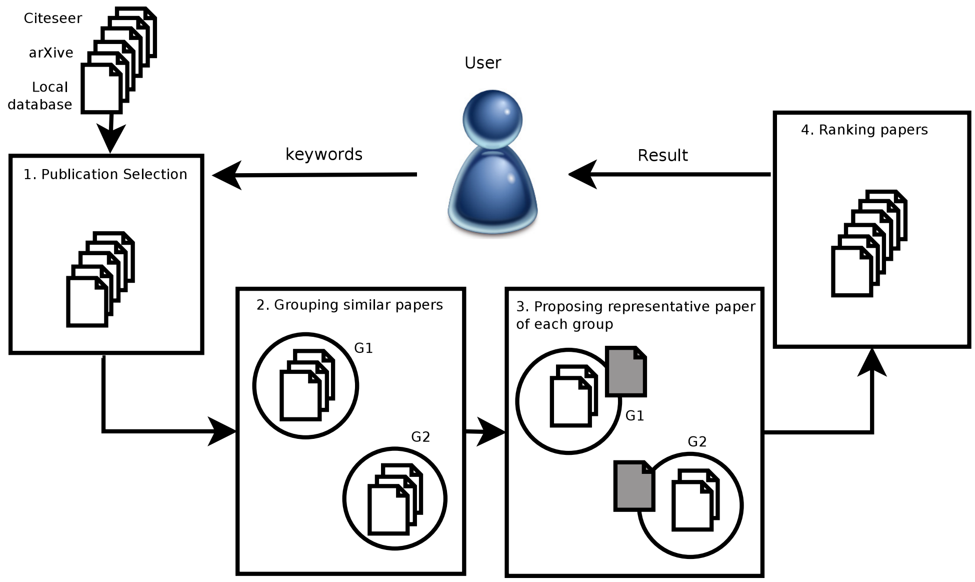

Figure 1. Method Overview

table IV which contains detail of each publication and table VI contains the list of citations from one paper to other ones.

B. Step 2: Grouping Papers

The second step consists in grouping the selected papers into clusters. Groups contain similar papers.

The Similarity σ between papers is computed by considering the titles, abstracts and common citations. We indeed assume that titles contain keywords, leading to the fact that if two titles share many common words, then this means they are similar. Moreover, we rely on the abstract as an indication of the content, thus assuming that common keywords lead to similar topics and interest.

Finally, as our approach aims at grouping papers that share common interest, we thus consider the co-citations.

These three criteria are aggregated using OWA so that it is possible to decide whether a representative paper is a paper being representative on all criteria or not.

Given two papers $d_1, d_2 \in D$, we thus have

$$\begin{aligned}\sigma(d_1, d_2) = \odot(&\sigma_K(T(d_1), T(d_2)),\\ &\sigma_K(Ab(d_1), Ab(d_2)),\\ &\sigma_C(C(d_1), C(d_2)))\end{aligned}$$

where:

- $\odot : [0,1]^n \to [0,1]$ is an aggregation operator for fusing the three similarity degrees, e.g., $\odot = OWA = avg, min, max, \ldots$
- $\sigma_K : \mathcal{P}(K)^2 \to [0,1]$ is a function comparing two sets of keywords and returning a number ranging from 0 to 1 which estimates to which extent the sets of keywords are similar;
- $\sigma_C : \mathcal{P}(D)^2 \to [0,1]$ is a function comparing two sets of cited papers and returning a number ranging from 0 to 1 estimating the similarity extent of the set.

As it is not relevant to consider that papers can be split into several groups in a crisp manner, we use fuzzy clustering, thus outputting overlapping groups. In this framework, we compute the membership degree of every paper p_i to every cluster c_j using the following equation:

$$u_{i,j} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\sigma(p_i - v_j)}{\sigma(p_i - v_k)} \right)^{\frac{2}{m-1}}}$$

C. Step 3: Electing Representative Papers

This step aims at proposing a representative paper of every group. A paper is considered as being representative if the topics are the ones that are shared in the group and if it has some criteria making it more interesting than other ones. For this purpose, the papers taken from a famous conference will be preferred to papers from non significant conferences.

Let c be a cluster containing the set of D_c papers, the representative paper $rep(c) \in D_c$ is computed as:

$$rep(c) = \arg \max_{\substack{p \in D_c \\ p' \in D_c \setminus \{p\}}} \sigma(p, p')$$

As we assume that it is not possible to find out only one paper being representative enough, we associate every representative paper to some other ones to complete the keywords that are not covered by the representative, as shown in Figure 2.

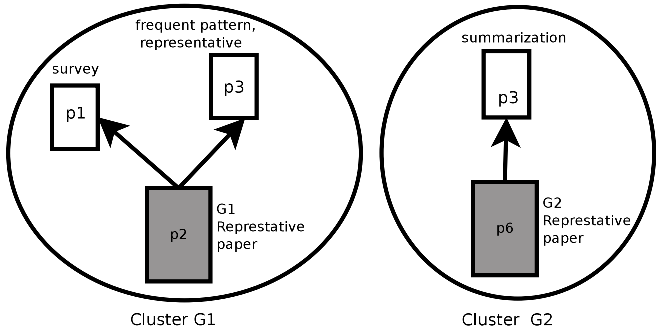

Figure 2. Representative Papers

D. Step 4: Ranking Representative Papers

The last step aims at ranking paper groups and reporting ordered papers to the user. For this purpose, we consider a measure such as the PageRank one.

VI. Conclusion

In this paper, we present our approach for paper recommendation. It relies on a workflow including soft approaches, thus allowing to take into account real dataset. It is indeed not relevant to consider crisp cuttings between papers and paper attributes.

Current and future works include the deep study of the measures used in our approach, for exploring efficiency for both semantic and computational (memory and time) criteria, together with a study of the evaluation process, for enhancing precision/recall criteria that are often used to assess the methods.

References

[1] E. Garfield and R. Merton, *Citation indexing: Its theory and application in science, technology, and humanities*. Wiley New York, 1979, vol. 8.

[2] M. Krapivin and M. Marchese, "Focused Page Rank in Scientific Papers Ranking." in *ICADL*, ser. Lecture Notes in Computer Science, G. Buchanan, M. Masoodian, and S. J. Cunningham, Eds., vol. 5362. Springer, 2008, pp. 144–153. [Online]. Available: http://dblp.uni-trier.de/db/conf/icadl/icadl2008.html#KrapivinM08;http://dx.doi.org/10.1007/978-3-540-89533-6_15;http://www.bibsonomy.org/bibtex/24724809764ff3ea95597cf97e7c62e90/dblp

[3] G. Adomavicius and A. Tuzhilin, "Towards the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 6, pp. 734–749, 2005.

[4] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry, "Using collaborative filtering to weave an information tapestry," *Communications of the ACM*, vol. 35, pp. 61–70, December 1992.

[5] M. Ekstrand, J. Riedl, and J. Konstan, *Collaborative Filtering Recommender Systems*. Now Publishers, 2011.

[6] R. Van Meteren and M. Van Someren, "Using content-based filtering for recommendation," in *Proceedings of the Machine Learning in the New Information Age: MLnet/ECML2000 Workshop*, 2000.

[7] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Analysis of recommendation algorithms for e-commerce," in *Proceedings of the 2nd ACM Conference on Electronic Commerce*. New York, NY, USA: ACM, 2000, pp. 158–167.

[8] R. Torres, S. M. Mcnee, M. Abel, J. A. Konstan, and J. Riedl, "Enhancing digital libraries with TechLens+," in *JCDL '04: Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*. New York, NY, USA: ACM Press, 2004, pp. 228–236.

[9] Z. Huang, W. Chung, T.-H. Ong, and H. Chen, "A Graph-based Recommender System for Digital Library," *JCDL '02*, pp. 65–73, 2002. [Online]. Available: http://delivery.acm.org/10.1145/550000/544231/p65-huang.pdf?key1=544231&%5C#38;http://www.bibsonomy.org/bibtex/21bdb59e9985512349189d5b41691fd55/brusilovsky

[10] M. Gori and A. Pucci, "Research paper recommender systems: A random-walk based approach," in *Web Intelligence, 2006. WI 2006. IEEE/WIC/ACM International Conference on*. IEEE, 2006, pp. 778–781.

[11] N. Agarwal, E. Haque, H. Liu, and L. Parsons, "Research paper recommender systems: A subspace clustering approach," *Advances in Web-Age Information Management*, pp. 475–491, 2005.

[12] C. Yang, B. Wei, J. Wu, Y. Zhang, and L. Zhang, "CARES: a ranking-oriented CADAL recommender system," in *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*. ACM, 2009, pp. 203–212.

[13] C. Shahabi and Y. Chen, "An adaptive recommendation system without explicit acquisition of user relevance feedback," *Distributed and Parallel Databases*, vol. 14, no. 2, pp. 173–192, 2003.

[14] Q. He, D. Kifer, J. Pei, P. Mitra, and C. L. Giles, "Citation recommendation without author supervision," in *Proceedings of the fourth ACM international conference on Web search and data mining*. New York, NY, USA: ACM, 2011, pp. 755–764.

[15] B. Gipp, J. Beel, and C. Hentschel, "Scienstein: A research paper recommender system," in *International Conference on Emerging Trends in Computing*, 2009, pp. 309–315.

[16] Y. Liang, Q. Li, and T. Qian, "Finding relevant papers based on citation relations," in *Proceedings of the 12th international conference on Web-age information management*, ser. WAIM'11. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 403–414.

[17] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *Journal of the American society for information science and technology*, vol. 58, no. 7, pp. 1019–1031, 2007. [Online]. Available: http://scholar.google.com/scholar.bib?q=info:U1ggap2AFkIJ:scholar.google.com/&output=citation&hl=en&as_sdt=0,5&as_vis=1&ct=citation&cd=0;http://www.bibsonomy.org/bibtex/234e3c294606a70cebdb0a2546d060b99/folke

[18] K. Sugiyama and M.-Y. Kan, "Scholarly paper recommendation via user's recent research interests." in *JCDL*, J. Hunter, C. Lagoze, C. L. Giles, and Y.-F. Li, Eds. ACM, 2010, pp. 29–38.

[19] V. Torra and Y. Narukawa, *Modeling decisions: information fusion and aggregation operators*. Springer-Verlag New York Inc, 2007.

[20] R. Yager, "On ordered weighted averaging aggregation operators in multicriteria decisionmaking," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 18, no. 1, pp. 183–190, 1988.

[21] R. R. Yager, "Families of OWA operators," *Fuzzy Sets Syst.*, vol. 59, no. 2, pp. 125–148, Oct. 1993. [Online]. Available: http://dx.doi.org/10.1016/0165-0114(93)90194-M

[22] J. C. Bezdek, "(Mostly) Fuzzy Clustering in Very Large Data Sets." in *IJCCI (ICFC-ICNC)*, J. Filipe and J. Kacprzyk, Eds. SciTePress, 2010, p. 5. [Online]. Available: http://dblp.uni-trier.de/db/conf/ijcci/ijcci2010-1.html#Bezdek10;http://www.bibsonomy.org/bibtex/2b518adfce1dbc041dabde674baacb152/dblp