FeRoSA: A Faceted Recommendation System for Scientific Articles

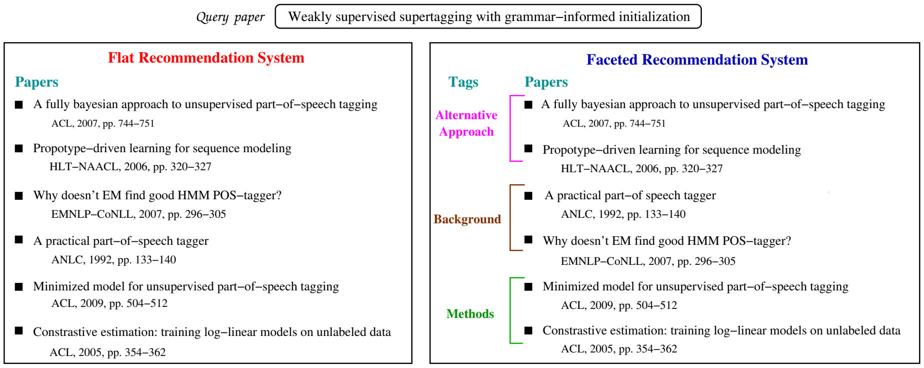
Tanmoy Chakraborty[1], Amrith Krishna[2], Mayank Singh[3], Niloy Ganguly[4], Pawan Goyal[5], and Animesh Mukherjee[6]

Department of CSE, Indian Institute of Technology, Kharagpur, India – 721302
{[1]its_tanmoy,[2]amrith.krishna,[3]mayank.singh}@cse.iitkgp.ernet.in
{[4]niloy,[5]pawang,[6]animeshm}@cse.iitkgp.ernet.in

Abstract. The overwhelming number of scientific articles over the years calls for smart automatic tools to facilitate the process of literature review. Here, we propose for the first time a framework of *faceted recommendation* for scientific articles (abbreviated as FeRoSA) which apart from ensuring quality retrieval of scientific articles for a query paper, also efficiently arranges the recommended papers into different facets (categories). Providing users with an interface which enables the filtering of recommendations across multiple facets can increase users' control over how the recommendation system behaves. FeRoSA is precisely built on a random walk based framework on an induced subnetwork consisting of nodes related to the query paper in terms of either citations or content similarity. Rigorous analysis based an experts' judgment shows that FeRoSA outperforms two baseline systems in terms of faceted recommendations (overall precision of 0.65). Further, we show that the faceted results of FeRoSA can be appropriately combined to design a better flat recommendation system as well. An experimental version of FeRoSA is publicly available at **www.ferosa.org** (receiving as many as 170 hits within the first 15 days of launch).

1 Introduction

One of the most common ways of doing any literature survey is perhaps the following – start from a known article and then traverse along those articles which have either cited the known article or have been cited by the known article. In particular, when a researcher reads the known article, she starts ruminating and asking recurrent questions pertaining to it that can further lead her to browse the other articles. These questions are most often synthesized based on the *knowledge context* of the users. For instance, an expert user, while reading a paper, might want to find papers presenting "alternative approach" of the query paper; while on the other hand, a naïve user might be interested to understand the "background" of the query paper. A smart recommendation engine should be able to organize the recommended papers into multiple such facets/tags. This would not only reduce the tedious effort of searching related articles, but also should answer a more fundamental question: what is the *role* of a recommended paper in relation to the query paper. However, the traditional paper recommendation systems primarily aim at improving the *relevance* of the recommendations and therefore tend to overlook the above fundamental aspect.


Query paper Weakly supervised supertagging with grammar−informed initialization

Flat Recommendation System

Papers

- A fully bayesian approach to unsupervised part−of−speech tagging
  ACL, 2007, pp. 744−751
- Propotype−driven learning for sequence modeling
  HLT−NAACL, 2006, pp. 320−327
- Why doesn't EM find good HMM POS−tagger?
  EMNLP−CoNLL, 2007, pp. 296−305
- A practical part−of−speech tagger
  ANLC, 1992, pp. 133−140
- Minimized model for unsupervised part−of−speech tagging
  ACL, 2009, pp. 504−512
- Constrastive estimation: training log−linear models on unlabeled data
  ACL, 2005, pp. 354−362

Faceted Recommendation System

Tags	Papers
Alternative Approach	A fully bayesian approach to unsupervised part−of−speech tagging ACL, 2007, pp. 744−751 Propotype−driven learning for sequence modeling HLT−NAACL, 2006, pp. 320−327
Background	A practical part−of speech tagger ANLC, 1992, pp. 133−140 Why doesn't EM find good HMM POS−tagger? EMNLP−CoNLL, 2007, pp. 296−305
Methods	Minimized model for unsupervised part−of−speech tagging ACL, 2009, pp. 504−512 Constrastive estimation: training log−linear models on unlabeled data ACL, 2005, pp. 354−362

Fig. 1: An illustrative example of a flat (left) and a faceted (right) recommendation system. The figure in the right side shows three facets (see Section 3) through which the recommended papers are related to the query paper.

In this paper, we attempt to build a "Faceted Recommendation System for Scientific Articles" (FeRoSA) that given a query paper, in addition to recommending the relevant scientific papers, organizes the recommendations into facets, thereby, suitably catering to the appropriate knowledge context of the end user. Our methodology is based on a principled framework of *random walk with restarts* that attempts to simulate the traversal mechanism of a user initiating from the known article. The model takes into consideration both the citation links as well as the content information to systematically produce the most relevant results. FeRoSA groups the recommendations into four naturally observed facets, namely, Background, Alternative Approaches, Methods and Comparison. This grouping has been formulated from the most intuitive forms of the knowledge context of the end users, which directly map to the different broad sections of any paper; however, the current system can very easily adapt to any other suitable form of grouping. A representative example of a flat and a faceted paper recommendation system is shown in Figure 1.

To the best of our knowledge, ours is the first *faceted* recommendation system for scientific articles. Moreover, the evaluation of such kind of systems is non-trivial and requires expert judgment. Therefore, we develop novel evaluation schemes. Due to the lack of standard baseline for faceted recommendation, we design two baseline systems and compare the performance of FeRoSA with these systems on the AAN (ACL Anthology Network) dataset [22]. The evaluation is conducted in three steps. First, the experts having significant domain knowledge are asked to develop the ground-truth dataset which is limited in size, based upon which initial evaluation of the faceted system is conducted. Second, a set of researchers having partial knowledge of the domain ("semi-experts") are asked for a mass-scale evaluation of the faceted system. Finally, we shortlist a few papers and request one of the authors of each paper to judge the quality of the recommendations returned by FeRoSA. We achieve an overall precision of 0.65 and 0.72 from the evaluation of experts and semi-experts respectively. As an additional objective, we further show that FeRoSA can also be appropriately modified to design a better flat recommendation system compared to Google Scholar, Microsoft Academic Search and a recent graph-based approach proposed by Liang et al. [16]. One of the possible reasons of success of the flat version of FeRoSA is the introduction of the diversity in the recommended articles; among the systems compared, FeRoSA seems to be the only one that returns recommendations which are both *relevant* and *diverse* at the same time.

2 Related work

Techniques applied for the traditional recommendation systems can be broadly classified into three categories: (i) *Collaborative filtering (CF)* [24], (ii) *Content-based methods* [13] and (iii) *Hybrid and other approaches* [1]. Several systems have also been developed particularly for scientific paper recommendation. Sugiyama and Kan [25] designed scholarly paper recommendation with citation and reference information. Lee et al. [15] proposed a personalized academic research paper recommendation system. Gipp et al. [8] developed "Scienstein" that can improve the approach of the typically used keyword-based search. Collaborative filtering in the domain of research paper recommendation has been criticized for various reasons [2].

In the context of designing search systems, there have been few attempts to incorporate facets [20]. Tunkelang [26] presented a novel approach that addresses the vocabulary problem for faceted data. Hearst [10] proposed a design guideline for faceted search interfaces. Bast and Weber [3] demonstrated the Semantic GrowBag approach to automatically organize facets for community-specific document collections. Recently, Diederich et al. [7] developed "FacetedDBLP" that allows to search computer science publications starting from some keyword and returns the result set along with a set of facets, e.g., distinguishing publication years, authors, or conferences. Sacco and Tzitzikas [23] presented theory and research results in dynamic taxonomy and faceted search systems. Vallet et al. [27] examined the use of multi-faceted recommendations to aid users while carrying out exploratory video retrieval tasks. In his master's thesis, Celma [4] studied music recommendation using a multi-faceted approach.

However, to the best of our knowledge, ours is the first attempt to introduce faceted paper recommendation system which unlike other real-time systems, is not only able to retrieve the relevant set of papers against a query paper, but also draws a semantic relation between the recommended papers and the query paper.

3 Dataset

We collected the AAN dataset[1] [22] which is an assemblage of all papers included in ACL[2] publication venues. In the full dataset, most of the papers had raw text. The texts were pre-processed where sentences, paragraphs and sections were properly separated using different markers. A significant part of the corpus had word splits and word joins. These were rectified using a dictionary based approach. The filtered dataset contains 9,843 papers (average 6.21 references per paper pointing to the papers within the dataset) and 7,892 unique authors.

We categorize the citation links based on their occurrence in various sections of the paper. Therefore, in addition to the citing-cited paper pair for each citation, we also

[1] **http://clair.eecs.umich.edu/aan/xml/**

[2] **https://www.aclweb.org/**

need to know the context and the section heading where the citation has occurred, in order to assign the facet. We use *Parscit* [6] to identify the citation contexts from the dataset and then extract the section headings for the pair of papers within the network.

Extraction of section heading. To extract the section heading, a list of 25,483 unique headings is collected and manually annotated into five different categories: Introduction, Related Work, Method, Results and Conclusion. The categories are further mapped into four facets, namely Background (Introduction), Alternative Approaches (Related Work), Method (Method) and Comparison (Results and Conclusion), as also suggested by Zhigang et al. [11]. A brief description of the facets/tags is as follows:

• **Background (BG)**: These are the citations which are prerequisite for understanding the basic notions of the citing paper. These citations generally point either to some seminal papers in that particular area, or to some papers which describe certain concepts that are relevant in understanding the framework of the citing paper. For instance, [9] is a suitable background paper for this study.

• **Alternative Approaches (AA)**: If there are citations to the approaches, which can be seen as alternative to the method proposed in the citing paper, then such citations are categorized as AA. These references are often found in system-oriented research papers where new methods/frameworks are proposed. For instance, for this paper, [16] may be treated as AA.

• **Methods (MD)**: If the citing paper borrows any such tools, techniques, datasets, measures or other concepts from the paper, or if both the papers have some overlap in usage of any of the entities mentioned above, then such a citation is treated as MD. For instance, [22] is a potential MD for this paper.

• **Comparison (CM)**: As mentioned in [16], a relation is said to be comparable if the citing paper has been compared to a cited paper in terms of differences or resemblances. Most of the times, these types of references tend to occur in the evaluation section of the citing paper. For instance, [16] is related to this paper by the CM tag. Essentially, one can argue that all the AA-tagged papers can be treated as CM papers. However, the AA-tagged papers may be irreproducible or difficult to be reimplemented, and thus may not be used for comparison. We only consider the cited paper as CM if it is used by the citing paper for comparison.

A facet is assigned to each pair of citing-cited paper, depending on the section information. Note that if a cited paper occurs multiple times in different sections of a citing paper, multiple facets would get assigned to this paper pair. Out of total 61,051 citation contexts extracted, the proportion in each facet is as follows: 23,022 (BG), 10,797 (AA), 8,828 (MD) and 18,404 (CM). To validate our approach of mapping section information to facets, we took experts' opinions[3] and obtained an average precision of 0.66. In parallel, we also performed an automated way of annotating references with the facets by Stanford MaxEnt classifier [17] with the features mentioned in [12], and

[3] The expert opinion was taken from the annotators, who were later involved in evaluating the systems as discussed in Section 5. For a direct reference of a paper, we asked experts whether the reference indicates BG, AA, MD or CM and then compared their opinion with our section annotation (in four categories).

obtained average precision of 0.68 (after 10-fold cross validation)[4]. We observed that the results obtained from the annotations using the section information and that from the supervised classification model were comparable. Moreover, the former is straightforward to compute and can be easily incorporated into a real application. Therefore, we proceed with the annotations obtained directly from the section information.

4 Recommendation method

In this section, we describe in detail the working principle of our proposed recommendation system. Figure 2 shows a schematic diagram of the proposed work-flow.

The citation network. We build the citation network which is a directed graph $G = (V, E)$ with edge labels. The labeling is a mapping from the edge set E to the set of facets based on the data obtained from the citation contexts. An edge may be tagged with multiple facets, if a paper cites another paper in multiple sections.

The induced subgraph. We construct an induced subgraph of the network for each query paper. An initial pool of vertices is obtained by following two criteria: (i) we consider all the papers which are at 1-hop or 2-hop distance from the query paper in the citation network irrespective of the label and directionality of edges; (ii) we also consider those papers that have a cosine similarity of at least 0.49 with the query paper (top 100 papers if the number of papers exceeds 100). Then we construct an induced subgraph of nodes present in the initial pool for each facet individually. For instance, for AA we only consider those citation edges in the induced subgraph which are labeled as AA. Note that in this process, few nodes might get disconnected or remain isolated. We connect these nodes with the query node through teleportation probability as discussed below.

Random walk on the induced subgraphs. One of the simplest ways to simulate the process of literature review based on knowledge context would be a suitable form of random walk that can mimic the article surfing behavior. Here, in order to obtain the importance of the nodes with respect to the query paper, we perform random walk with restarts (RWR) [19] on the induced subgraph with query paper being the starting node. RWR is defined in Equation 1: consider a random walker that starts the walk from node i. The walker iteratively moves to its neighborhood with a probability proportional to the edge weights. At each step of the random walk, it has some probability c to return to the starting node i. The relevance score of node j with respect to node i is defined by the steady-state probability $r_{i,j}$ that the walker will finally stay at node j:

$$\overrightarrow{r}_i = (1-c)\hat{A}\overrightarrow{r_i} + c\overrightarrow{e}_i \tag{1}$$

where $\overrightarrow{r}_i = [r_{i,j}]$ is an $n \times 1$ ranking vector; $r_{i,j}$ is the relevance score of node j with respect to node i; c is the restart probability, $0 \leq c \leq 1$ (we consider $c = 0.4$ [19]); $\hat{A}$ is the normalized weighted matrix associated with the weighted adjacency matrix $A = [a_{ij}]$; $\overrightarrow{e}_i$ is the restart vector, with all its elements 0 except the i^{th} element.

Apart from the citation links in the induced subgraph, we also consider the isolated nodes by assigning a *teleportation probability* (i.e., a probability of randomly jumping

[4] In the interest of space, the detailed experiments and results of the supervised classification are not presented in this paper.

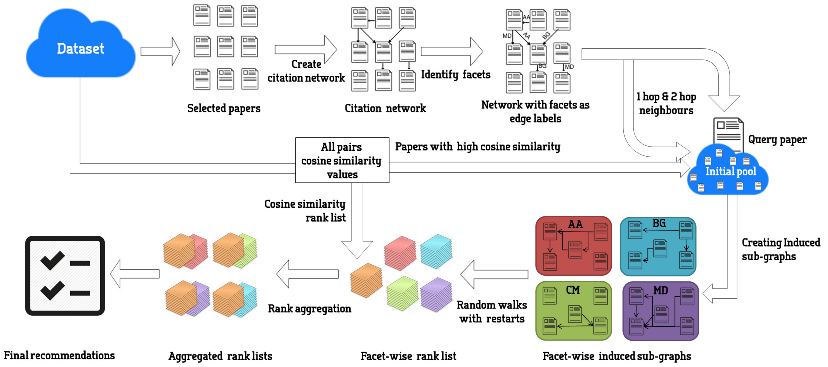


Fig. 2: (Color online) The work-flow diagram of `FeRoSA`.

to any one of the isolated nodes) as 0.3, thus eliminating the chance of the isolated nodes remaining unreachable by the random walker.

Rank aggregation. We use the above framework to obtain a rank list of nodes present in the induced subgraph for each facet separately. Additionally, we consider content similarity by measuring the cosine-similarity between the query paper and each of the papers present in the induced subgraph. Next, we utilize a rank aggregation method to combine these two types of rankings. In our work, we use *RankAggreg* [21], where the rank aggregation is considered as an optimization problem to find an ordered list δ that minimizes the total distance between each of the provided lists L_i and δ. Note that for each facet T, we aggregate the ranking obtained for T and the cosine-similarity based ranking to obtain the final rank list. In addition, we also perform a total rank aggregation in order to design a flat version of `FeRoSA` (`f-FeRoSA`) by combining all the facet-wise rankings and the cosine similarity based ranking together (see Section 5.3).

Design principles. All the sub-tasks involved in `FeRoSA`, such as sub-graph creation, edge labeling, RWR and rank aggregation, can be performed independently for each query paper. To find out the recommendations for the entries, we do not require the entire (global) network to be in-memory; rather we require only the 2-hop neighbors of each query paper, and the relevant top k documents based on pre-calculated cosine similarity values. Therefore, we anticipate that the system will work within similar time bound per query, irrespective of the size of the dataset. Moreover, we noticed that on an average the entire process for a single query takes 382 ms, with the largest one being 497 ms. When new entries need to be inserted into the existing system, we need not recalculate recommendations for the entire dataset, but only for those entries which are affected by the new entries. Hence, `FeRoSA` is scalable and light-weight.

5 Experimental Results

In this section, we present the performance of `FeRoSA`. We design a new evaluation framework, consisting of three independent steps: (i) experts' judgment on a limited set of papers, (ii) semi-experts' judgment for mass-scale evaluation, and (iii) the judgment by the original authors of the paper. Finally, we show that `FeRoSA` can also be used to design a better flat recommendation system.

5.1 Evaluation metrics

We use *Overall Precision* (OP) and *Overall Impression* (OI) for comparative evaluation. OP measures the ratio between the number of relevant recommendations (according to the experts' judgments) and the total number of recommendations provided for a query paper by each competing system. The OP of each system is then measured by averaging OPs for all the query papers. OI measures that among all the query papers, in how many cases a particular system is rated to have an overall better performance. We measure this value for a system on the basis of precision majority. Similarly for faceted evaluation, we measure the OP of the recommended papers under each individual facet.

5.2 Evaluation of faceted recommendation

In this section, we first describe the process of ground-truth generation, followed by a brief description of the baseline algorithms, and then elaborate the comparative evaluation of all the systems for faceted recommendation.

Ground-truth generation. Because of the unavailability of a benchmark dataset for the evaluation of scientific article recommendation especially for the faceted recommendation, we conducted an expert judgment to generate a set of faceted and flat recommendations (used later in Section 5.3) as our ground-truth. First, we shortlisted a set of 30 query papers that cover the fields of expertise of 10 experts. For each query paper, we presented 30 recommendations that we pulled from four separate systems: FeRoSA, Google Scholar (GS)[5], Microsoft Academic Search (MAS)[6] and a graph based paper recommendation system proposed by Liang et al. [16] (LLQ henceforth). Note that the three latter systems which are quite popular for paper recommendation are further used in Section 5.3 as competing systems to FeRoSA for flat recommendation. The experts were provided with web based interfaces[7], in which they were shown 30 recommendations for each query paper (the name of the systems remained anonymous). Each expert had to mark whether each recommended paper was relevant to the query paper, and if so, the possible facet(s).

Baseline systems. Due to the lack of faceted recommendation system for scientific literature, we design two competitive baseline systems to compare with FeRoSA.

• **VanillaPR:** For each query paper, we form a single induced subgraph $G'(V', E')$ which is exactly the same as that built for FeRoSA but ignores the facet labeling. Once the graph is formed, we perform RWR from the query paper. We then retrieve the nodes having the highest values from RWR. Finally, to label the retrieved papers with facets, we train a supervised model using the ground-truth data we collected in the experts' judgment. We use the following three types of features to train the supervised model. For each pair of the query and the recommended paper, we use total 12 (4 Boolean + 8 real valued) features: (i) section of the recommended paper, if it appears in 1-hop of the query paper (4 Boolean features, one for each section), (ii) within V' for the query paper, fractional number of times a particular recommended paper appears in a given

[5] **http://scholar.google.co.in**

[6] **http://academic.research.microsoft.com/**

[7] **http://www.ferosa.org/evaluation/**

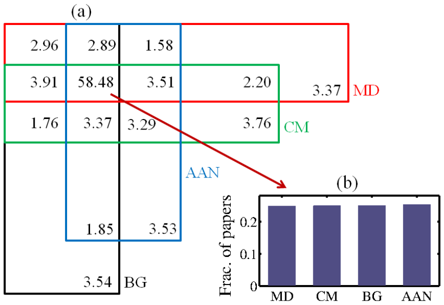


Fig. 3: (Color online) (a) Venn diagram of the recommended papers in four facets; (b) distribution of papers which are tagged by all four facets.

Table 1: Faceted evaluation of all the competing faceted recommendation systems. For VanillaPR, the results are reported by performing 3-fold cross validation.

Facets	VanillaPR	FeRoSA-CS	FeRoSA
BG	0.65	0.51	**0.79**
AA	0.48	0.34	**0.56**
MD	**0.62**	0.39	**0.62**
CM	0.44	0.38	**0.62**
Average	0.55	0.40	**0.65**

section (4 real-valued features, one for each section), and (iii) for a given recommended paper, fraction of times it is cited in a given section by any paper in the whole dataset (4 real-valued features, one for each section). We then learn the weights for features with a rankSVM model [14]. We report the average precision of the system after performing a three fold cross-validation over the ground-truth data.

• **FeRoSA-CS:** Our second baseline recommends papers by relying only on RWR, performed on subgraphs of papers within 2-hop distance from the query paper, without considering the cosine-similarity (CS) based papers. This in turn answers the necessity of considering cosine-similarity based papers while constructing the initial pool.

Comparative analysis. We conduct an empirical study on the results obtained from FeRoSA. Figure 3(a) represents a Venn diagram of all the recommended papers under different facets, i.e., in what facets have a particular paper been recommended for different queries. For example, 3.54% of the recommended papers appear only as BG to any of the query paper. We observe that 58.48% of the recommended papers appear under all the four facets for various queries. For these papers, we further show in Figure 3(b) their distribution in different facets, which seems to be fairly uniform.

We report in Table 1 the OP of all the systems for different facets. We observe that FeRoSA attains the highest OP (0.65) amongst all other systems, which is 18% and 62.5% higher than VanillaPR and FeRoSA-CS respectively. The maximum OP of FeRoSA is obtained for BG (0.79), which is followed by MD (0.62), CM (0.62), and AA (0.56). The pattern is also similar for FeRoSA-CS. For the case of MD, however, we observe similar performance for VanillaPR and FeRoSA.

For further analysis, we present the confusion matrix for the facets in Table 2 along with the false positive rate (FPR) for each system. The FPR is quite low for the facets (i.e., the specificity values of the facets are quite high). To understand the reason behind the misclassification, we unfold the tagged citation network once again and observe that it arises due to the frequent occurrences of a single edge being tagged by multiple facets. For instance, Table 2 shows that AA is mostly misclassified to MD for FeRoSA. In the tagged citation network, we observe the same phenomenon that among the multi-faceted edges with AA tag, around 50% of edges are tagged by both AA and MD (similarly for CM and MD with 36.6% of occurrences together).

Facets	AA	BG	CM	MD	FPR
AA	29	11	15	8	0.08
BG	13	87	11	17	0.09
CM	7	9	19	11	0.06
MD	11	6	7	32	0.05

(a) VanillaPR

Facets	AA	BG	CM	MD	FPR
AA	23	14	9	17	0.09
BG	18	69	13	28	0.14
CM	13	8	18	7	0.06
MD	11	14	5	26	0.05

(b) FeRoSA-CS

Facets	AA	BG	CM	MD	FPR
AA	33	11	6	13	0.07
BG	14	99	5	10	0.07
CM	2	6	28	10	0.04
MD	3	7	12	34	0.04

(c) FeRoSA

Table 2: Confusion matrix for the faceted evaluation (false positive rate: FPR). For VanillaPR, we report the confusion matrix for that run where maximum OP for all the facets is achieved.

Mass-scale evaluation. To broaden the evaluation of FeRoSA, we perform a mass-scale evaluation, aiming for more coverage on the system output and targeting a wider set of evaluators. All the selected evaluators had a good knowledge of the NLP domain. This time we reverse engineer the process by selecting few papers from the ground-truth data, each of which appears in the recommendation of multiple query papers. To start with, we shortlisted a collection of 31 such recommended papers. For each recommended paper, we enlisted the set of query papers (and the facets) in which the recommended paper has appeared[8]. The evaluators then evaluated the relevance of the recommended paper, as well as the relevance of the facet with respect to each query paper in which the given recommended paper has appeared. Total 26 experts participated in this evaluation task. For each recommended paper, we calculate OP per facet for all its corresponding query papers and show the results in Table 3(a). Similarly, we calculate facet-wise OP for all the query papers and report it in Table 3(b). We observe that for both the cases, FeRoSA significantly outperforms other baselines.

(a)

Facets	VanillaPR	FeRoSA-CS	FeRoSA
BG	0.57	0.70	**0.73**
AA	0.43	0.41	**0.53**
CM	0.37	0.59	**0.64**
MD	0.58	0.55	**0.69**
Avg.	0.49	0.56	**0.64**

(b)

Facets	VanillaPR	FeRoSA-CS	FeRoSA
BG	0.66	0.82	**0.85**
AA	0.45	0.48	**0.54**
CM	0.42	0.54	**0.73**
MD	0.51	0.68	**0.77**
Avg.	0.51	0.63	**0.72**

Table 3: Overall precision per facet for (a) recommended paper to query paper and (b) query paper to recommended paper.

As a related objective we investigate whether our system performs better for the highly-cited query papers, or whether the same accuracy is achieved for all citation ranges of the query papers. Generally, a standard recommendation system should perform equally well for all ranges of query papers [15]. Here we divide the entire range of incoming citations of the query paper into three buckets and measure the facet-wise OP of all the competing systems for each bucket separately. In Table 4, we see that FeRoSA performs significantly better than the other baseline systems even for low-cited query pa-

[8] This indeed reduced the evaluators' effort of reading multiple papers.

Facets	VanillaPR			FeRoSA-CS			FeRoSA		
	Low	Medium	High	Low	Medium	High	Low	Medium	High
BG	0.53	0.73	0.71	0.44	0.54	0.55	0.65	0.84	0.87
AA	0.41	0.52	0.49	0.28	0.35	0.41	0.53	0.56	0.61
MD	0.57	0.59	0.71	0.40	0.34	0.44	0.65	0.55	0.67
CM	0.29	0.55	0.48	0.33	0.39	0.41	0.56	0.62	0.69
Avg.	0.45	0.59	0.59	0.36	0.40	0.45	**0.59**	**0.64**	**0.71**

Table 4: Performance of three competing faceted systems for different query papers divided into three citation ranges (Low: 0 to 6, Medium: 7 to 28, High: 29 to 343).

pers.

Evaluation by the authors. There is no better alternative than the authors themselves when it comes to evaluating the recommendation for a particular paper. We were curious to know whether FeRoSA could impress the authors with its recommendations. We therefore designed a judgment experiment by selecting a set of 30 authors and sent each of them, a judgment form, where we specified one of his/her papers as query paper, and one (top) recommendation from FeRoSA for each facet. The author had to make a binary judgment about the relevance of recommendation to the query as well as the relevance of the facet for the recommendation separately. A sample response of an anonymous author can be found at **http://ferosa.org/#authorresponse**. Twelve authors responded to the survey. We obtain an average precision of 0.50 (BG: 0.49, AA: 0.42, MD: 0.52, CM: 0.59). In 75% cases the recommended papers are marked as relevant. Four authors marked three out of four faceted recommendations as relevant. Overall, the authors appreciated the attempt of designing a faceted recommendation system for scientific articles.

5.3 Evaluation of flat recommendation

We further posit that FeRoSA can also be used as a flat recommendation system if the rank lists obtained from the different facets and the cosine-similarity based ranking can be appropriately combined. Therefore, we use the rank-aggregation method discussed in Section 4 in order to obtain a flat recommendation list.

In this section, we discuss the performance of the flat version of FeRoSA (f-FeRoSA henceforth) and compare it with three state-of-the-art flat baseline systems: Google Scholar (GS), Microsoft Academic Search (MAS) and a graph based paper recommendation system, LLQ [16]. We consider LLQ as a baseline system because similar to our approach, it also classifies citation relations into three categories, namely Based-on, Comparable and General using the approach proposed in [18], and these categories are further used to compute a final combined score. Note that while GS and MAS are mostly known for searching scientific papers, an inherent nature of ranking of the retrieved results has lead us in using them as potential baseline systems.

We perform a broad analysis of the performance of all the competing methods. Table 5(a) reports the values of individual metrics mentioned earlier, averaged over all the judgments conducted by the experts. For top three recommendations per system, f-FeRoSA achieves OP of 0.79 which is 29%, 75% and 62% higher than GS, MAS and LLQ respectively. One can also notice that for 43% of the cases, f-FeRoSA fares

(a)

System	Relevance		Diversity		
	OI@3	OP@3	ISP	σ_1	σ_2
GS	0.27	0.61	0.043	24.92	66.29
MAS	0.17	0.45	0.043	16.51	53.85
LLQ	0.13	0.41	0.054	21.43	63.82
f-FeRoSA	**0.43**	**0.79**	**0.003**	**54.77**	**108.93**

(b)

OP	f-FeRoSA
OP@3	0.79
OP@5	0.78
OP@10	0.71

Table 5: (a) Flat evaluation of the competing systems based on relevance and diversity; (b) overall precision of f-FeRoSA at different number of recommendations.

better than all other systems in terms of OI. Clearly, f-FeRoSA is preferred nearly twice more than GS, which is the second best performing system. This indeed shows that f-FeRoSA outperforms the state-of-the-art recommendation systems by a reasonable margin. We also see in Table 5(b) that f-FeRoSA is quite consistent in recommending highly relevant papers within top rank list.

As mentioned earlier, the reason behind the success of f-FeRoSA can be the introduction of *diversity*, i.e., inclusion of highly relevant papers from each facet into the aggregated list. To substantiate this argument quantitatively, we further take two graph-based measures from [5] and evaluate the competing flat systems based on diversity: (i) *Inverse Shortest Path* (ISP)=$\frac{\sum_{u,v\in S} 1/d_{uv}}{|S|\times(|S|-1)}$, (ii) *Expansion Ratio* (σ_l)=$\frac{\bigcup_{u\in S} N_u^l}{|S|}$, where S: set of recommended papers, d_{uv}: shortest path between u and v in the citation network, and N_u^l: neighbors within l-hops of u in the citation network (we take $l = 1, 2$). The less (more) the ISP (σ_l), the more the diversity. Results in Table 5(a) corroborate our argument that f-FeRoSA is indeed the most diverse system among others.

6 Discussions and Future work

In this paper, we proposed a faceted recommendation system, FeRoSA for scientific articles that not only recommends relevant articles for a particular query paper, but also relates the recommended articles with the query paper through four pre-defined facets. As a by-product of this study, we also obtain an annotated citation network where each citation link between a citing paper and a cited paper gets labeled, and a ground-truth dataset for evaluating scientific recommendation systems. FeRoSA is designed to be light-weight, so that it can easily be deployed as an online system. We evaluated our system in three stages based on human judgment and observed significant performance improvement. Although FeRoSA is designed for faceted recommendation, we further showed that it significantly outperforms the baselines in flat recommendation.

The scalability of our proposed model can be guaranteed after its extensive usage and testing on other datasets. We are also interested in the design aspects related to the ergonomics of the user interface so that it can significantly reduce user's cognitive overload, while providing high user satisfaction at the same time. We anticipate that the framework used in FeRoSA can be adopted to design faceted recommendations for items such as movies, books, videos etc. The annotated citation network and the human evaluation results are available at **www.ferosa.org/data**.

References

1. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE TKDE 17(6), 734–749 (Jun 2005)

2. Agarwal, N., Haque, E., Liu, H., Parsons, L.: Research paper recommender systems: A subspace clustering approach. In: WAIM. LNCS, vol. 3739, pp. 475–491 (2005)
3. Bast, H., Weber, I.: Type less, find more: fast autocompletion search with a succinct index. In: SIGIR. pp. 364–371. ACM, Seattle, Washington (2006)
4. Celma, Ò.: Music Recommendation: A multi-faceted approach. Master's thesis (2006)
5. Chakraborty, T., Modani, N., Narayanam, R., Nagar, S.: Discern: A diversified citation recommendation system for scientific queries. In: ICDE. pp. 555–566. Seoul (2015)
6. Councill, I.G., Giles, C.L., Kan, M.Y.: Parscit: an open-source crf reference string parsing package. In: LREC. Marrakech, Morocco (2008)
7. Diederich, J., Balke, W.T., Thaden, U.: Demonstrating the semantic growbag: Automatically creating topic facets for faceteddblp. In: JCDL. pp. 505–505. ACM, New York, USA (2007)
8. Gipp, B., Beel, J., Hentschel, C.: Scienstein: A research paper recommender system. In: ICETiC. pp. 309–315. Virudhunagar, India (2009)
9. He, Q., Pei, J., Kifer, D., Mitra, P., Giles, L.: Context-aware citation recommendation. In: WWW. pp. 421–430. ACM, New York, NY, USA (2010)
10. Hearst, M.A.: Design recommendations for hierarchical faceted search interfaces. In: Proc. SIGIR 2006, Workshop on Faceted Search. pp. 26–30. Seattle, Washington (2006)
11. Hu, Z., Chen, C., Liu, Z.: Where are citations located in the body of scientific articles? a study of the distributions of citation locations. Journal of Informetrics 7(4), 887–896 (2013)
12. Jochim, C., Schütze, H.: Towards a generic and flexible citation classifier based on a faceted classification scheme. In: COLING. pp. 1343–1358. Mumbai, India (2012)
13. Koren, Y.: Factorization meets the neighborhood: A multifaceted collaborative filtering model. In: SIGKDD. pp. 426–434. New York, USA (2008)
14. Lee, C.P., Lin, C.J.: Large-scale linear ranksvm. Neural computation 26(4), 781–817 (2014)
15. Lee, J., Lee, K., Kim, J.G.: Personalized academic research paper recommendation system. CoRR abs/1304.5457 (2013)
16. Liang, Y., Li, Q., Qian, T.: Finding relevant papers based on citation relations. In: Web-Age Information Management, vol. 6897, pp. 403–414. Springer (2011)
17. Manning, C., Klein, D.: Optimization, maxent models, and conditional estimation without magic. In: NAACL. pp. 8–8. ACL, Stroudsburg, USA (2003)
18. Nanba, H., Okumura, M.: Towards multi-paper summarization using reference information. In: IJCAI. pp. 926–931. Morgan Kaufmann, Stockholm, Sweden (1999)
19. Pan, J.Y., Yang, H.J., Faloutsos, C., Duygulu, P.: Automatic multimedia cross-modal correlation discovery. In: SIGKDD. pp. 653–658. Seattle, WA (2004)
20. Peng, T.C., cho Timothy Chou, S.: itrustu: a blog recommender system based on multifaceted trust and collaborative filtering. In: SAC. pp. 1278–1285. ACM, Hawaii, USA (2009)
21. Pihur, V., Datta, S., Datta, S.: Rankaggreg, an r package for weighted rank aggregation. BMC bioinformatics 10(1), 62 (2009)
22. Radev, D., Muthukrishnan, P., Qazvinian, V., Abu-Jbara, A.: The acl anthology network corpus. LREC pp. 1–26 (2013)
23. Sacco, G.M., Tzitzikas, Y.: Dynamic Taxonomies and Faceted Search: Theory, Practice, and Experience. Springer Publishing Company, Incorporated, 1st edn. (2009)
24. Sarwar, B., Karypis, G., Konstan, J., Riedl, J.: Item-based collaborative filtering recommendation algorithms. In: WWW. pp. 285–295. ACM, New York, NY, USA (2001)
25. Sugiyama, K., Kan, M.Y.: Scholarly paper recommendation via user's recent research interests. In: JCDL. pp. 29–38. ACM, New York, NY, USA (2010)
26. Tunkelang, D.: Dynamic Category Sets: An Approach for Faceted Search. In: ACM SIGIR. ACM New York (2006)
27. Vallet, D., Halvey, M., Hannah, D., Jose, J.M.: A multi faceted recommendation approach for explorative video retrieval tasks. In: IUI. pp. 389–392. New York, USA (2010)