Evaluation of Contextualization and Diversification Approaches in Aggregated Search

Hermann Ziak
Know-Center GmbH
Inffeldgasse 13
8010 Graz, Austria
hziak@know-center.at

Roman Kern
Know-Center GmbH
Inffeldgasse 13
8010 Graz, Austria
rkern@know-center.at

***Abstract*—The combination of different knowledge bases in the field of information retrieval is called federated or aggregated search. It has several benefits over single source retrieval but poses some challenges as well. This work focuses on the challenge of result aggregation; especially in a setting where the final result list should include a certain degree of diversity and serendipity. Both concepts have been shown to have an impact on how user perceive an information retrieval system. In particular, we want to assess if common procedures for result list aggregation can be utilized to introduce diversity and serendipity. Furthermore, we study whether a blocking or interleaving for result aggregation yields better results.**

In a cross vertical aggregated search the so-called verticals could be news, multimedia content or text. Block ranking is one approach to combine such heterogeneous result. It relies on the idea that these verticals are combined into a single result list as blocks of several adjacent items. An alternative approach for this is interleaving. Here the verticals are blended into one result list on an item by item basis, i.e. adjacent items in the result list may come from different verticals.

To generate the diverse and serendipitous results we relied on a query reformulation technique which we showed to be beneficial to generate diversified results in previous work. To conduct this evaluation we created a dedicated dataset. This dataset served as a basis for three different evaluation settings on a crowd sourcing platform, with over 300 participants. Our results show that query based diversification can be adapted to generate serendipitous results in a similar manner. Further, we discovered that both approaches, interleaving and block ranking, appear to be beneficial to introduce diversity and serendipity. Though it seems that queries either benefit from one approach or the other but not from both.

INTRODUCTION

Cross Vertical Aggregated Search in Information Retrieval is defined as the unification of several, often distributed, data or knowledge bases into one search interface. This setting provides a number of benefits but also bears a lot of challenges. The most obvious problem is the presentation of the different results from the verticals to the user, which might be very heterogeneous, e.g. news, pictures, video, text. Literature refers primarily to three main challenges: collection representation, collection selection and result merging. [1], [2] Within this work we focus on the topic of result merging. In particular, on the challenge of how to best introduce a certain degree of diversity and serendipity into the final result. [3] These two concepts have been shown to have an impact on the suitability of the information for users in information retrieval.

The presented work emerged from the EEXCESS [1] (Enhancing Europes eXchange in Cultural Educational and Scientific reSources) project. The goal of this project is to recommend high quality content to users from a large range of different knowledge sources from the field of cultural heritage and scientific literature. The project is open-source and can be obtained from GitHub[2]. In our scenario users are assumed to be in the state of a creative process, e.g. researching a certain topic or writing a document. The user gets recommendations, which are based on their current context and their past behavior. Unlike a traditional federated search system the query is not explicitly stated by the user but automatically inferred by the user context detection component. [4] Such a scenario is also known as Just-in-Time Information Retrieval. [5] According to literature, this automatic process might lead to an underrepresentation of the true information need of the users. Our work is motivated to counter-steer this tendency by introducing the concepts of serendipity and diversity into the result list generation.

The main question is how to achieve a certain amount of diversity and serendipity within an aggregated item list. Furthermore, the question arises which already well studied techniques from the field of federated search could be incorporated.

The evaluation of such a mixed approach, i.e. covering precise results but with a certain degree of novelty, is a challenge for a number of reasons. Many of the most common methods and measures, which are discussed in section II, to evaluate such a mixed approach, does not seem to fit. Furthermore, there is a lack of ground truth datasets with which new approaches can be compared against each other. Because of these reasons we decided to approach the evaluation by the use of crowd sourcing. This option is becoming more and more popular in recent years and gives us the opportunity to estimate the suitability of the approach in the absence of such ground truth data. Therefore we conducted several sub-evaluations on

[1] http://eexcess.eu/
[2] https://github.com/EEXCESS/recommender.git

the crowd sourcing platform CrowdFlower[3]. Over 300 workers participated producing more than 1500 judgments. Within this work we distinguish between the people taking part in the evaluation, referred as the workers, and the people that are supposed to interact with such a system, referred to as the users.

Related Work

Even though recommender systems are often measured by metrics covering accuracy [6], literature suggests that highly precise results might not be that useful in some cases [7]. For example, when items are proposed purely based on their relevancy by a recommender system it is likely that are for the most part already known by the users [8]. Therefore, methods to introduce a broader, yet helpful, list of recommended items is needed. To achieve such a result, diversity and serendipity might play an important role. Diversity could be described as resolving the ambiguity of a query [9]. This is usually accomplished by first capturing the different meanings of a query and then producing results that respond to each of the meanings individually. The result of the result diversification is a list of items that should all be relevant to the query, but are dissimilar to each other. In contrast to diversity, Toms [10] describes serendipitous information retrieval as the occurrence of a user interacting with an information node with no prior intentions to do so. Thus results are generated that are not strongly related to the query but associated with other aspects of a user's background, in order to create a result list, that can be best described by "pleasantly surprising". While it is trivial to generate results that are artificially highly diverse and have the potential to be serendipitous, the results are still required to prove to be beneficial to the users.

In the case of diversity, there are two potential starting points to deal with this problem. On one hand the query itself can be altered to generate diversified results [11]. On the other hand, algorithms such as IA-Select [3] exist, where the result list is altered to be intend aware. The first kind of method has the potential to be adopted to generate results that are at the same time also serendipitous.

The core of most of the federated search system techniques lies in the result list aggregation. Typically there are three main approaches how results from different sources (or verticals) can be aggregated: *Non Blended*, *Blended* [12] and most recently *Composite Retrieval* [13]. The *Non Blended* approach presents results from each source in a separated view. This concept is often applied for verticals like news or videos on major search engines. In contrary, the blended integration mixes results from different resources into one single list. *Composite Retrieval* is somehow a fusion of both approaches were bundles of topically related documents are returned. Out of this three possibilities the *Blended* approach seems to be best suited to introduce diversity and serendipity. Literature suggest several distinct techniques on how a blended approach can be realized. We selected two of these approaches to compete against each other: interleaving [14] and block ranking [15]. The interleaving approach mixes different verticals into the original result list on certain positions. In contrary the blocking approach combines the verticals as blocks, without blending them, to generate the final result list.

Traditionally the Cranfield paradigm [16] is followed to evaluate approaches in information retrieval. Measures like mean average precision (MAP) or normalized discounted cumulative gain (NDCG) can be used to evaluate the performance in an offline manner. Unfortunately, these approaches do not appear to be useful for result list that introduces diversity and serendipity. We assume that such methods would be punished by an evaluation that is strictly focused on relevancy, underestimating their potential to be helpful. In particular within a scenario where the users do not explicitly state their information need such queries that are purely navigational are expected to play a minor role. Diversity measures like NDCG intent aware (NDCG-IA) [3] and α-NDCG [9] do exist, but they do not cover serendipity and probably won't be feasible to be applied on blended result lists. In particular, the generation of a ground truth data specifically for serendipity remains an open issue. Therefore a method to evaluate the usefulness of our methods without resorting to a ground truth data-set is needed. We opted for a method, which has gained popularity recently - namely crowd sourcing [17].

Approach and Evaluation Design

Dataset Creation

To be able to conduct a crowd sourcing based evaluation we had to decide whether we allow the workers to specify their own queries, or to prepare a set of predefined queries. In order to obtain comparable results we decided to prepare a set of queries before the actual evaluation. As first step queries were selected from query logs out of the EEXCESS project itself [18]. The final set of query consisted of 52 different queries.

Diversity was achieved by expanding the initial query via pseudo relevance feedback based on a knowledge base. This approach is described and discussed in previous work in detail [19], [20]. To introduce serendipity into the result list the user's history of visited paged was analyzed as well. These pages were aggregated and condensed. Out of this condensed pages, the main topics and terms that represent the users past context were extracted. This approach is related to the diversity approach since the query is expanded as well. The principal difference is within the process of formulating the query. Instead of creating a disjunction query with artificially added terms, it is formulated as conjunction query where the expanded terms represent the history or interests of the user. This is expected to create a query drift [21] that should lead to serendipitous results. See Figure 1 as example for both approaches.

As source we opted to query an existing knowledge base, in our case the English version of Wikipedia. This decision was based on the idea that a data set containing multiple sources could have made it difficult to generalize our findings.

[3]https://www.crowdflower.com/

A detailed description of this index can be found in previous work [20].

The set of queries together with the generated previous interests were passed through the system to create a total of three different result lists representing the three different modes of our system: i) The basic result lists that had no diverse and no serendipitous results included, ii) The diversified result list, iii) The result list that contained the potential serendipitous results.

For the evaluation these lists were used to populate the final result lists according to the desired result aggregate method being studied:

- **Basic** The list users were told to compare with as baseline containing the items of the unaltered results.
- **Interleaved** A list where top diverse and serendipitous results were interleaved into the basic result list.
- **Blocked** The list containing three blocks: A block of basic results, as second the top diverse results (de-duplicated against the basic results) and as third block the serendipitous results (again de-duplicated against previous results).
- **Diverse** Similar to the blocked list, but this time only containing basic results and diversified results.
- **Serendipitous** A list consisting of just two blocks, the first being populated from the basic results and the second block containing the serendipitous results.

Evaluation Scenarios

In order to conduct our evaluation we decided to have the workers compare two different search results, being displayed side-by-side. The workers than have to vote for the list being more in line with their expectations. In addition, the task description contains the criteria to watch out for. These criteria differ in the varying evaluation scenarios.

Out of these 5 different configurations (basic, interleaved, blocked, diverse and serendipitous) a total of three evaluation scenarios emerged:

a) Evaluation Scenario #1: The first evaluation scenario contained either the basic and the block list or the basic and the interleaved list. The purpose of this evaluation was to get a basic understanding of the acceptance level of potential users for such approaches. The workers were introduced to get into the mindset of a potential user and were given additional information about the query and the user's history. In this set-up the algorithms introducing diversity and serendipity could both potentially bias the overall results of the evaluation. Thus the second evaluation scenario was created.

b) Evaluation Scenario #2: Here the worker had to decide against the shortened basic list and the lists containing either a diversity block or a serendipity block. With this evaluation we wanted to address the possibility that one of the approaches has a severe adverse effect on the other.

c) Evaluation Scenario #3: The third and last scenario was a direct comparison of the block ranking approach and the interleaved approach. If it may prove to be too difficult for the workers to get into the mindset of the potential user then this set-up should help to draw conclusions about the performance of the blocking and interleaved approach.

To get an adequate amount of ratings for each query each task was conducted six times by different workers. To reduce the potential risk of a bias towards the list presented at first each set of tasks was split into two sub sets where the lists where mirrored against each other.

Within the crowd sourcing system each task contained further information on how the workers have to perform the task. For example, to get the workers into the mindset of the users that carried out the original query, the extracted terms of the respective user's history used to generate the serendipitous list were summarized in a short sentence. We opted for this summarization over the whole list, as pre-test with friendly users signaled that they were overwhelmed when showing all terms.

Once the workers have stated their preference to one of the presented lists, they were also asked to state how hard it was to render their final decision.

The crowd sourcing platform we used gives the workers the possibility to provide feedback and rate certain characteristics of the task. Here the workers stated, with an average of 4.4 out of 5, that the instruction was clear and unambiguous for the evaluations. As an additional measure to prevent workers from giving fluke ratings we added the limitation that the result was flagged if the worker answered within less than half a minute.

To summarise, these are the questions that we tried to answer by conducting the evaluation:

- How does the block ranking perform against the basic result list?
- How does the interleaved list perform against the basic result list?
- How do both approaches perform compared directly against each other?
- Can serendipitous results be achieved by query reformulation on a similar level than diversification?

Evaluation Limitations

Commonly used measure like Fleiss' κ [22] for inter-rater agreement are not applicable to this evaluation since the overlap of users rating the same set of queries is limited. Therefore we decided to report the agreement on item level with the arithmetic mean of the percentage of the biggest agreement. See Equation 1 for illustration. Where S is the set of all tasks within the evaluation, a_i and b_i are the sums of votes towards one of the algorithms with the restriction that a_i is always bigger or equal b_i.

$$f(x) = \frac{\sum_{i \in S} \left(\frac{a_i}{a_i + b_i} | a_i \geq b_i \right)}{|S|} \tag{1}$$

The last reported figures are the percentage of the selected algorithm per approach according to the preference of the workers.

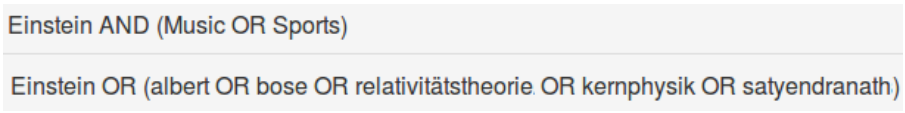

Fig. 1. This figure shows two queries send to the system. The top row shows a serendipitous query. The expanded terms form a disjunction with each other and are conjunct with the original query terms to produce a query drift. The lower row shows a diversified query where the block is also disjunct to boost intents that might be underrepresented.

TABLE I
RESULTS FOR THE INTERLEAVING AND BLOCKING APPROACH. REPORTED VALUES ARE RESULTS OF THE ARITHMETIC MEAN AGREEMENT ON ITEM LEVEL AND THE DECISION PERCENTAGE TOWARDS THE STATED APPROACH WITHIN THE CORRESPONDING ROW.

	Item Agreement	Decision Percentage
Interleaved	0.692	0.358
Blocked	0.721	0.355

TABLE II
RESULTS FOR THE SHORTER BLOCKED LISTS CONTAINING EITHER ONLY DIVERSIFIED AND BASIC RESULTS OR SERENDIPITOUS AND BASIC RESULTS. BOTH APPROACH HAVE SIMILAR AGREEMENTS AND DECISION PERCENTAGE.

	Item Agreement	Decision Percentage
Diverse	0.769	0.307
Serendipitous	0.746	0.31

RESULTS

Table I represents the results for the first evaluation scenario. The agreement is similar at about 70 percent for both algorithms. We further analyzed the queries that got more votes for either the blocked or the interleaved approach. Here only one query was present in both datasets that did not receive the majority of the votes for both approaches.

The results in Table II represent the second evaluation scenario. The goal of this scenario is a comparison of both approaches to ensure that one has no strong adverse effects upon the other.

Table III shows the results of the third evaluation scenario where the block ranking approach is directly compared against the interleaving approach. The item agreement of the users is about 65 percent. Both approaches obtain similar results with a slight tendency towards the block ranking approach.

The level of confidence for each of the tasks from the workers is reported in Table IV. The results of the first scenario, referred as ”Interleaved” and ”Blocked” in the table, are very similar. Less than 20 percent of all iterations of all task were labeled as hard to decide. Whereas over 30 percent were labeled as very confident. In the second scenario, referred as ”Serendipitous” and ”Diverse”, the workers seemed to be more confident than within the other scenarios. Nearly half of the workers stated that they were very confident with they answer and on in about 15 percent of the cases the workers were unsure. The third and last scenario, referred as ’Blocked vs Interleaved’, shows that the amount of very confident answers is the lowest in all the runs.

TABLE III
RESULTS FOR THE INTERLEAVING APPROACH COMPARED DIRECTLY AGAINST THE BLOCKING APPROACH. BOTH APPROACHES OBTAIN SIMILAR RESULT, THOUGH A SLIGHT TENDENCY TOWARDS THE BLOCK RANKING APPROACH SEEMS TO EXIST.

	Item Agreement	Decision Percentage
Blocked vs Interleaved	0.647	0.532

TABLE IV
RESULTS OF THE FEEDBACK OF THE USERS HOW CONFIDENT THEY WERE WITH THEIR DECISION BETWEEN THE PAIR OF LISTS IN PERCENT. EACH USER HAD TO STATE HIS CONFIDENCE FOR EACH TASK.

	Interleaved	Blocked	Seren.	Diverse	Blocked vs Inter.
Very Conf.	30.0	34.0	47.5	42.5	27.0
Confident	52.0	47.0	38.0	41.5	59.0
Hard	18.0	19.0	14.5	16.0	18.0

DISCUSSION

Within scenario #1 we wanted to assess the general acceptance level of the interleaved and the blocking approach. Given that the user were most probably not familiar with the topics, the assessment did not cover their personal information need and the history did not reflect their own experience we consider the acceptance rate as sufficient. The provided link for the keywords to Web search engine was used several times for each query, even though a short explanation of the topic was already provided. This can be seen as an indicator that some users were not familiar with the topic of the query. Both approaches yield similar results but a closer inspection of which query benefited the most from which setting revealed that only six percent of queries overlapped. This leads to the conclusion that both algorithms are justifiable and yield different user satisfaction. This is also evidenced by the confidence values, presented in Table IV. Here the amount of ”very confident” answers is the lowest compared to all others. Thus our recommendation is that if one wants to implement such a system is advised to investigate into a learn to query approach and use both approaches alternating depending on the underlying query.

Within scenario #2, where the main question was whether the diversity or serendipity approach had adverse effects on each other, we could show that the acceptance rate and agreements are on similar levels. We also assume that the query formulation process to generate the serendipitous results works as well as the diversification approach since the ”Item Agreement” and ”Decision Percentage” as well as the confidence statements are on similar levels.

The results of scenario #3 show as well that the users had more difficulties to decide which approach produced more helpful results. Here a tendency for higher percentages towards the block ranking approach can be seen. Although it has to be considered that this could just be based on the fact that the first block was presenting more elements of the unaltered list in comparison to the first elements of the interleaved list.

Conclusion and Future Work

In general, we conclude that our approach to introduce diversity and serendipity appear to work on similar levels. Compared to the baseline list the according to the decision percentage, both approaches, block ranking and interleaving, do not consistently out-perform the unaltered result list. Although that can be partly explained by the fact that workers might have had problems putting themselves into the position of the original user, highlighted by the tendency of workers to gather more information about the queries using Web search engines. This also constitutes one of the findings of our study, namely taking the knowledge and background of the worker within a crowd sourcing platform into account. The direct comparison of the two aggregation approaches showed that some queries seem to benefit from the blocking approach while others seem to benefit from the interleaving approach. The two set appear to be disjunct. Therefore we plan to investigate this further and consider to re-evaluate our dataset using learn to query techniques to decide which approach to apply for each query.

Acknowledgments

The presented work was developed within the EEXCESS project funded by the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement number 600601. The Know-Center is funded within the Austrian COMET Program - Competence Centers for Excellent Technologies - under the auspices of the Austrian Federal Ministry of Transport, Innovation and Technology, the Austrian Federal Ministry of Economy, Family and Youth and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG.

References

[1] M. Shokouhi and L. Si, "Federated search," *Foundations and Trends in Information Retrieval*, vol. 5, no. 1, pp. 1–102, 2011.

[2] J. Lu and J. Callan, "Federated search of text-based digital libraries in hierarchical peer-to-peer networks," in *Advances in Information Retrieval*. Springer, 2005, pp. 52–66.

[3] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong, "Diversifying search results," in *Proceedings of the Second ACM International Conference on Web Search and Data Mining*. ACM, 2009, pp. 5–14.

[4] J. Schlötterer, C. Seifert, and M. Granitzer, "Web-based -in-time retrieval for cultural content," *PATCH14: Proceedings of the 7th International ACMWorkshop on Personalized Access to Cultural Heritage*, 2014.

[5] B. J. Rhodes, "Just-in-time information retrieval," Ph.D. dissertation, Massachusetts Institute of Technology, 2000.

[6] T. Murakami, K. Mori, and R. Orihara, "Metrics for evaluating the serendipity of recommendation lists," in *New frontiers in artificial intelligence*. Springer, 2007, pp. 40–46.

[7] S. M. McNee, J. Riedl, and J. A. Konstan, "Being accurate is not enough: how accuracy metrics have hurt recommender systems," in *CHI'06 extended abstracts on Human factors in computing systems*. ACM, 2006, pp. 1097–1101.

[8] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl, "Evaluating collaborative filtering recommender systems," *ACM Transactions on Information Systems (TOIS)*, vol. 22, no. 1, pp. 5–53, 2004.

[9] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon, "Novelty and diversity in information retrieval evaluation," in *Proceedings of the 31st annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2008, pp. 659–666.

[10] E. G. Toms, "Serendipitous information retrieval." in *DELOS Workshop: Information Seeking, Searching and Querying in Digital Libraries*. Zurich, 2000.

[11] R. L. Santos, C. Macdonald, and I. Ounis, "Exploiting query reformulations for web search result diversification," in *Proceedings of the 19th International Conference on World Wide Web*. ACM, 2010, pp. 881–890.

[12] S. Sushmita, H. Joho, M. Lalmas, and R. Villa, "Factors affecting click-through behavior in aggregated search interfaces," in *Proceedings of the 19th ACM International Conference on Information and knowledge management*. ACM, 2010, pp. 519–528.

[13] H. Bota, K. Zhou, J. M. Jose, and M. Lalmas, "Composite retrieval of heterogeneous web search," in *Proceedings of the 23rd International Conference on World Wide Web*. ACM, 2014, pp. 119–130.

[14] A. Chuklin, A. Schuth, K. Hofmann, P. Serdyukov, and M. de Rijke, "Evaluating aggregated search using interleaving," in *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*. ACM, 2013, pp. 669–678.

[15] J. Arguello, F. Diaz, and J. Callan, "Learning to aggregate vertical results into web search results," in *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*. ACM, 2011, pp. 201–210.

[16] E. M. Voorhees, "The philosophy of information retrieval evaluation," in *Evaluation of cross-language Information Retrieval Systems*. Springer, 2002, pp. 355–370.

[17] O. Alonso, D. E. Rose, and B. Stewart, "Crowdsourcing for relevance evaluation," in *ACM SigIR Forum*, vol. 42, no. 2. ACM, 2008, pp. 9–15.

[18] C. Seifert, J. Schlötterer, and M. Granitzer, "Towards a feature-rich data set for personalized access to long-tail content," in *Proceedings of the 30th Annual ACM Symposium on Applied Computing*. ACM, 2015, pp. 1031–1038.

[19] H. Ziak and R. Kern, "Evaluation of pseudo relevance feedback techniques for cross vertical aggregated search," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Springer, 2015, pp. 91–102.

[20] R. Rubien, H. Ziak, and R. Kern, "Efficient Search Result Diversification via Query Expansion Using Knowledge Bases," in *Proceedings of 12th International Workshop on Text-based Information Retrieval (TIR)*, 2015.

[21] A. M. Lam-Adesina and G. J. Jones, "Applying summarization techniques for term selection in relevance feedback," in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2001, pp. 1–9.

[22] J. L. Fleiss, "Measuring nominal scale agreement among many raters." *Psychological bulletin*, vol. 76, no. 5, p. 378, 1971.