Topic Assisted Fusion to Re-Rank Texts for Multi-Faceted Information Retrieval

Rajendra Prasath[1], Aidan Duane[2] and Philip O'Reilly[1]

[1] University College Cork (UCC), Cork, Ireland
[2] Waterford Institute of Technology, Waterford, Ireland
{R.Prasath, Philip.OReilly}@ucc.ie; aduane@wit.ie

Abstract. We propose to develop a framework for an intelligent business information system with multi-faceted data analysis capabilities that supports complex decision making processes. Reasoning and Learning of contextual factors from texts of financial services data are core aspects of the proposed framework. As part of the proposed framework, we present an approach for the ordering of contextual information from textual data with the help of latent topics identified from the web corpus. The web corpus is prepared by specifically using a number of financial services sources on the web that describe various aspects of mobile payments and services. The proposed approach first performs weighting of query terms and retrieves the initial set of texts from the web corpus. We use Latent Dirichlet Allocation (LDA) on this web corpus to identify the topics that relate to the contextual features of various financial services/products. The retrieved texts are scored based on the identified topics that could cover a variety of contextual factors. We performed subjective evaluation to identify the relevance of the contextual information retrieved, and found that the proposed approach captures a variety of key contexts pertaining to user information needs in a better way with the support of topic assisted contextual factors.

1 Introduction

The commercial importance for companies in generating effective business models is emphasised by Teece [1], who states that adopting the correct business model when offering new products and services is critical for organisational performance management and enterprise success. In order to adopt and implement an appropriate business model, it is necessary for organisations to have an in-depth knowledge of the existing market and to learn from successful product/service implementations which the organisation already has in the market[1]. Following a review of the patent library, it is evident that there is no commercially available decision support system that enables organisations to research the specifics of the market in which they operate, in order to enable them, to review the marketplace and to develop, test and validate suitable business models for their products/services.

With the advent of the semantic web and the great amount of business data available online, this research focuses on transforming the highly interlinked documents of the world wide web into a rich knowledge base[2]. The primary focus is on developing an intelligent business information system with multi-faceted data analysis for better decision making in relation to business models. In this paper, we present an approach based on topic assisted information fusion to perform a re-ranking of texts retrieved for a given information need. The preliminary analysis shows that the proposed algorithm records a much greater depth and quality of results vis a vis the baseline.

2 Motivation

In the business model domain, the state of the art is very descriptive. Osterwalder's paper, based on the business model canvas[3], is typically used as an application by which practitioners can analyse and construct business models for their business ecosystem. While this framework is acknowledged within industry as being the "state of the art", an analysis of same reveals a number of key limitations like poor predictive capacity, no correlated evaluation of factors associated with business ecosystem, inability to combine internal and external knowledge base on a specific product or service. Many organisations struggle to get external information pertaining to current market trends, innovations and competitors in an efficient way. Many managers within organisations utilise commercial search engines like Google, Bing, etc to retrieve information and manually navigate through them. This is an inefficient, time consuming approach and is typically not fused with an organisations' internal knowledge base.

There are interesting fusion based reasoning approaches in the literature. Chang *et* al. [4] proposed an interactive reasoner abbreviated as *Pequliar* that applies progressive query language and interactive reasoning (cum learning) for information fusion support. However, this research is limited in that the reasoner is guided by humans to elaborate the query by means of some rule and then a query processor uses this to produce a more informative answer. Park and Kim [5] proposed an interactive grey-zone case based reasoning model that makes decisions focusing additional attention on cases near cut-off point. This work emphasizes organisations' need to learn from previous cases and experiences, especially in relation to designing and commercialising new products and services. Especially each organization is supposed to have internal knowledge about their products, services, partners and customers. External sources would have information illustrating the impacts of their products, voices of their customers and business critics in the market. By fusing an organisations' internal knowledge with that from external sources, organisations would have a greater insight on the market and be in a position to learn and apply similar contexts to solve current similar problems which they face. In order to assist organisations to capture external knowledge efficiently, we have derived a framework that uses topic assisted information fusion to capture similar contexts, and then apply these contexts to learn to solve similar instances from a knowledge base in an non-interactive way. Subsequently we would fuse this external knowledge with the internal knowledge of the interested companies involved in mobile payments sector. In this paper, the term “context” means the semantic association between the specific user query and the retrieved set of texts. The terms, *aspect* or *facet*, are used interchangably to represent the type of contextual relations between the query and the retrieved set of texts.

3 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a generative probabilistic topic model [6, 7]. To analyse a discrete collection of data, especially text corpora, LDA applies three-level hierarchical Bayesian model, in which each item of a collection is modelled as a finite mixture over an underlying set of topics. Each topic is, in turn, modelled as an infinite mixture over an underlying set of topic probabilities. In the context of text modelling, the topic probabilities provide an explicit representation of a document. LDA defines the marginal distribution of a document as a continuous mixture distribution, as follows,

$$p_i(x) = p(d|\alpha_i, \beta_i) = \int p(\theta|\alpha) \left(\prod_{j=1}^{n} p(w_j|\theta, \beta) \right) d\theta \tag{1}$$

where d is a document with n words. $p(\theta|\alpha)$ and $p(w_j|\theta, \beta)$ are actually multinomial distributions with Dirichlet prior. $p(w_j|\theta, \beta)$ describes $K-$dimensional topic-word distributions. The parameters α, β are estimated by means of Gibbs sampling [8]. Here α is the symmetric Dirichlet prior for all documents and β is the symmetric Dirichlet prior for all topics

4 Topic Assisted Information Fusion

Reasoning about data with lots of dynamics is vital for many applications where the basic data sources come from different types, and the generated data is heterogeneous. Data obtained from such sources are often ambiguous and associated with certain levels of uncertainty. Topic models could narrow down the search to specific areas by reducing the level of uncertainty to a greater extent. The use of topic models in information retrieval is well studied in the literature [9]. In this work, we use data from multiple web sources and perform information fusion assisted by the information on latent topics to improve the quality of text retrieval. The retrieved texts provide market and organisational knowledge to the user, enabling better decision making. The original query is expanded based on these aspects that a set of more informative answers to the specific users information needs is retrieved. We plan to capture the aspects of the query in terms of distinct topics covered by it. Then using these topics associated information, we score the retrieved texts and re-rank them to bring the informative content to the top. Then the reasoner accepts the top ranked texts and identifies the matching of similar texts/cases that could potentially represent similar contexts. The learner is enforced to work on these text segments to perform context sensitive assessments so as to update both the learning experience and the knowledge base.

4.1 The System Design

The proposed framework, as shown in Figure. 1, consists of three major components: *Query Processing*, *Topic Assisted Information Fusion,* and *Knowledge Management Process*

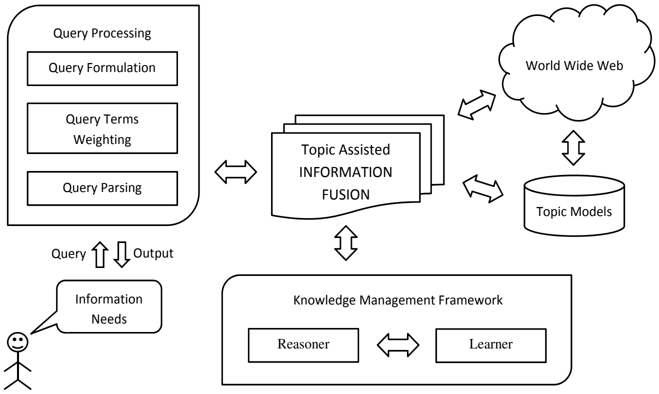

Fig. 1. The proposed retrieval framework with reasoner and learner

Query Processing: The user enters their information needs in terms of a set of keywords or key phrases or short sentences typically consisting of 3-5 keywords. Bai and Nie [10] used query expansion using term relations based on language models that integrates several contextual factors to adapt specific query contexts. The proposed system obtains the query terms and processes them to find the associated aspects and then the query term weighting is applied to identify their individual weights depending on the corpus statistics. The weighting of a query term q_i is computed as follows:

$$qtw(q_i) = averageTF(q_i) * IDF(q_i), \forall i \in [1, n] \tag{2}$$

where $averageTF(q_i)$ is computed as the ratio between the total number of occurrences of the given query term across the texts in the collection and the total number of texts (in terms, the total number of texts in which the term occurs) and $IDF(q_i) = \log(\frac{N}{df(q_i)})$ where N = total number of texts and $df(q_i)$ = frequency of texts given the query term q_i.

The proposed query term weighting technique focuses on important query terms that may represent an entity (the primary focus), rather than other associated query terms of this entity. For example, consider the query: *countries adopting mobile payments*. In this query, “countries” is the main focus and it might represent the number of countries adopting mobile payments, or the name of the countries adopting mobile payments or the type of payment services in countries adopting mobile payments or the status of the countries offering mobile payments. We obtain the following weight for each of these query terms [weights in brackets] as follows: countries[6.3057], adapting[4.2886], mobile[1.0530], payments[2.0595].

Consider another query: *How will mobile payments integrate coupons?* In this query, customer loyalty services such as special offers, coupons, bonus points for either using their service or buying their products are the main focus. This reflects in the weight estimated for the individual term: mobile[1.4307], payments[1.7373], integrate[2.8587], coupons[2.9401]. These weights are used as the boost factors in the formulation of the expanded query. In this paper, we considered the system employing this weighting approach as the *baseline* system for the retrieval of texts. Here "text" means a meaningful sentence or passage having qualitative and quantitative information pertaining to the user query.

Topic Assisted Information Fusion: This component first seeks the interpretation of the query with latent topic models and infers the user information need in terms of the coverage of a variety of topics learned from the given corpus. It then uses the expanded query to retrieve information from the index and applies fusion on each text retrieved with the topic associated information to compute the contextual similarity score. The contextual similarity score of each retrieved text is computed as follows:

$$tscore(td_i) = s * \sum_{j=1}^{m} qtw(w_j) + (1-s) * \frac{1}{|td_i|} \sum_{j=1}^{m} \frac{1}{nt} * \sum_{l=l}^{p} prob(w_j|T_l) \qquad (3)$$

where $td_i = \{w_1, w_2, \cdots, w_m\}$, $s = cosine(query, retrieved\ text)$, nt denotes the number of topics in which the term is associated with and $|td_i|$ denotes the number of unique words in the retrieved text. Here the topic model probability contributes the likelihood score of a term given an aspect and cosine similarity score contributes the degree of relevance of the retrieved text given a query. So the combined *tscore* is used to represent a variety of aspects that are relevant to the user needs. Then the retrieved texts are ranked in decreasing order of their contextual similarity. This produces a set of top ranked texts containing key information pertaining to the user's query. This task is different from cluster based approaches, as described in [11], in which is the query is matched against the cluster of documents instead of individual documents and clusters are ranked based on their similarity to the query.

Knowledge Management (KM) Process: This consists of two tasks: *Reasoning* and *Learning*. The proposed KM framework is supposed to process the top ranked list of documents and identify similar text segments from the index. This task is yet to mature for reporting. Actually we plan to find the patterns of similar text fragments, with high matching scores, that are sent to the learner with the actual user query to perform context sensitive assessments. The knowledge gained by the learner could be reused to solve similar scenarios as guided by Case Based Reasoning(CBR) [12, 13].

4.2 Proposed Approach

The proposed system works as follows: A user enters their information needs in the form of a query - a sequence of keywords. The system receives this query and applies the proposed query terms weighting approach using corpus statistics to understand the level of importance attached to these query terms, in order to capture the actual context of the user's query. Using the weight computed for each query term, the system re-formulates the given query into the weighted query. This weighted query is used to retrieve texts from the search system. The initial set of results retrieved by the search system is assumed to be relevant, as similar to the Pseudo-Relevance Feedback (PRF) approach. We use the standard cosine similarity as the scoring function during the texts retrieval. The retrieved results are fused with topic assisted information and re-ranked based on their contextual similarity. Then the re-ranked results are evaluated subjectively to compute the retrieval efficiency of the top 20 results. This research is evolving towards the goal of developing a more sophisticated fusion approach that combines the results obtained with multiple sources.

The pseudo code of the proposed approach is illustrated in Algorithm 1.

Algorithm 1 Topic Assisted Information Fusion to Re-rank Texts

```
Require: A Searcher - gets user query and retrieves top $k$ texts $D = \{td_1, td_2, \cdots, td_k\}$ where $td_i = \{w_1, w_2, \cdots, w_m\}$
Input: Query $Q$ having a sequence of keywords: $\{q_1, q_2, \cdots, q_n\}$
Description:
  Input: Enter the user query into the system
  Query Terms Weighting: Extract the query terms and use corpus statistics to determine their individual weights using Equation. 2.
  Text Retrieval: Compute $s = cosine(Q, D)$ and retrieve top $k$ texts
  Topic Assisted Information Fusion:
  Create the topic model with $p$ topics (fixed) by applying LDA on the entire corpus
  for all $td_i \in D$ do
      Compute weights of each term in $td_i$, $qtw(w_j)$ and take their summation.
      Compute $tscore(td_i)$ using Equation. 3
      Update Texts with the computed $tscore(td_i)$
  end for
  Re-Rank: Choose top $k$ texts based on high $tscore(td_i)$ scores
  return top $k$ texts ($k \leq n$)
Output: The ranked list of top $k \leq n$ texts pertaining to the query context
```

5 Experimental Results

5.1 Corpus

We have created a collection of web documents / reports crawled from the websites of various financial companies using open source web crawlers. We collected 14,764 web documents containing a total of 296,983 words. Actually we do not have the idea about the coverage of documents pertaining to mobile payments sector in the TREC web collections. So we have chosen to crawl much more focused information pertaining to mobile payments. To achieve this, we have selected 20 contextually different types of user information needs pertaining to mobile payments and created the web corpus. These queries are chosen as the representative sample of the statistical population that comes from the documents in the underlying web corpus. However, the test of significance has to be done to justify this choice. We used the open source implementation of JGibbLDA[1] for building the topic models. For this, we are limited to 200 latent topics with 1,000 iterations to build the topic model (assumed other LDA parameters: $\alpha = 0.5$ and $\beta = 0.1$).

The selected queries that pertain to the specific information needs in the mobile payment sector are listed in Table. 1. Most of the queries contain the phrase, "mobile payments" due to the fact that the underlying information needs are very much specific to the issues in mobile payments. Even though the above listed queries represent questions and look like seeking answers, as in the Q & A systems, the primary focus is on retrieving texts whose context matches the actual context of the given query. By representing the "context" of the query, we mean different aspects pertaining to the focused information need of the given query. For example, consider query - Q4 (Table. 1) which searches for mobile payment models. The retrieved texts should contain the details about the payment models currently available in the market. Business models of payment services would also be considered as the relevant context.

QID	**Actual Information Needs (Queries)**
Q1	What are mobile payments?
Q2	How will mobile payments benefit consumers?
Q3	How will mobile payments benefit retailers?
Q4	What is the cost to consumers of using mobile payments?
Q5	What is the cost to retailers of mobile payments?
Q6	How is the mobile payments security addressed?
Q7	How is the mobile payments privacy addressed?
Q8	What technology do retailers need to accept mobile payments?
Q9	How will mobile payments integrate coupons?
Q10	What challenges arise with mobile payments?
Q11	How are mobile payments protected by legislation?
Q12	What types of mobile payments are available?
Q13	How big is the global mobile payments market?
Q14	How mature is the global mobile payments market?
Q15	What is the most common type of mobile payment?
Q16	What is the most common value of mobile payments?
Q17	What companies are the biggest players in the global mobile payments market?
Q18	Can mobile payments be hacked?
Q19	Will mobile payments replace cash?
Q20	Which retailers accept mobile payments?

Table 1. List of Queries

[1] http://jgibblda.sourceforge.net

5.1.1 Evaluation Methodology

We have used the following steps for the subjective evaluation to test the quality of the content retrieved:

- The focus is on evaluating the quality or goodness of the document content in terms of the coverage of informative subtopics pertaining to the query.
- We have used 2 evaluators to judge the quality of the content in the top 20 results retrieved two systems: the standard *vector space model*(VSM) [14] with the proposed query weighting approach as the baseline system and the system with the proposed re-ranking approach.
- The evaluator reviewed one query at a time and the top 20 ranked documents retrieved for that query.
- For each query, each evaluator picked up top 20 ranked texts and evaluated them by analysing their context. The number of facets covered by each text pertaining to the query is identified. Then, depending on the variety of facets and their importance pertaining to the query, the quality of the content is scored in the 3 point scale as outlined in Table. 2.
- Final scores are used to compute $p@d$ for both lists.

It is not a fair idea to consider VSM as the baseline method rather than other standard PRF methods like Rocchio or probabilisitc methods because the primary focus is on incorporating information pertaining to latent topics in payments sector. The following scale is used for subjective evaluation:

Score	Description
1.0	various aspects of the query context
0.5	partial information of the query context
0	NOISY / irrelevant information

Table 2. Guidelies for the subjective evaluation

While evaluating the pieces of information retrieved by the base line and the proposed systems, the evaluators are instructed to focus on the relevance of the texts retrieved with respect to the query context. We do not apply deep NLP parsing on the textual content (except sentence level parsing with '.', '?', '!' as sentence markers). But in this work, we try to get PROBABLE Text fragments that could represent the expected context of the user information needs(especially financial service payment oriented queries).

We applied the following evaluation measure: *Precision* at top d texts (in short, $p@d$) to evaluate the ranked list of top d (= 5, 10, 20) results retrieved for each query. We used the three different tiny datasets for this experiment with the same set of queries listed in Table. 1. The top 5, 10 and 20 results were manually evaluated and the observations relating to the retrieved texts are discussed in the subsequent section.

6 Discussion

In this section, we present our key observations on the nature of the texts retrieved for the specific information needs. During the analysis, we have considered the top 5, 10 and 20 texts retrieved for each query and analysed the context represented by the texts. The context of the texts retrieved is analysed to find its matching with the context of the query. Figure. 2 shows the text retrieval performance of the baseline and the proposed approach with "Precision @ top 5" (p@5) scores. We present some of our key observations in analysing the top 5 texts retrieved. For query Q1, the primary focus of the user is to find information pertaining to the cost to be paid by the customers to use mobile payment services. The baseline results have only one partially relevant result whereas the proposed system fetched texts having the information on the cost imposed by the payment service providers to access their mobile payment services. Similar observations were observed for query Q4, in finding the cost involved in using mobile payment services and for Q5, in identifying the cost-benefit trends of retailers of mobile payments sector. The best performance is achieved for query Q10, which identifies the distinct types of mobile payments activity observed in the market. For this query, we retrieved texts having the details of various messaging services like text messaging, simple message services (SMS), etc. Also we have identified qualitative information having the details of various electronic accessaries involved in mobile payments sector. We also observed a better performance in query Q9, in which the security related aspects of mobile payments are well addressed in the retrieved texts and the query Q12, in which benefits of the retailers are well addressed with the information on their low and high margins.

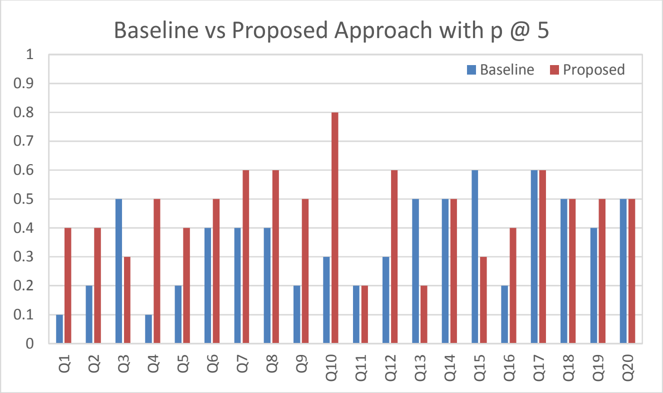

Fig. 2. p@ 5 comparison: Baseline vs Proposed retrieval approach

Query drifting takes place for specific queries Q3, Q13 and Q15 with the topic assisted approach. We analyse these queries individually. The query Q3 focuses on the information about the coupons offered by mobile payments service providers to attract their customers. The retrieved texts constitute more context-insensitive collections of information than are associated with the statistics of loyalty programs in mobile payments sector. For query Q13, the query is intended to find the possible pitfalls in hacking of mobile payments. But the retrieved texts contain descriptions on hacked histories pertaining to various companies and phones involved in mobile sector. “7500 mobile phones had been hacked live in China” and “hacking of SquareUp mobile payment system” are some observations in the retrieved texts. For query Q15, the actual intent is to find the size of the global mobile payments market. For this query, noisy texts, that are not filtered during content extraction, result in the decrease of performance. However the retrieval performance improved with the subsequent 15 texts retrieved.

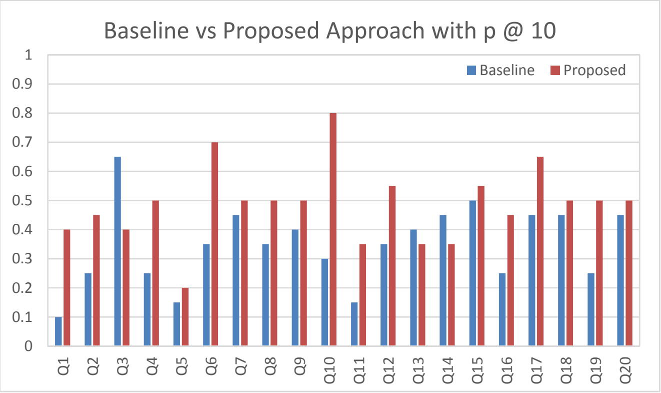

Fig. 3. *p*@ 10 comparison: Baseline vs Proposed retrieval approach

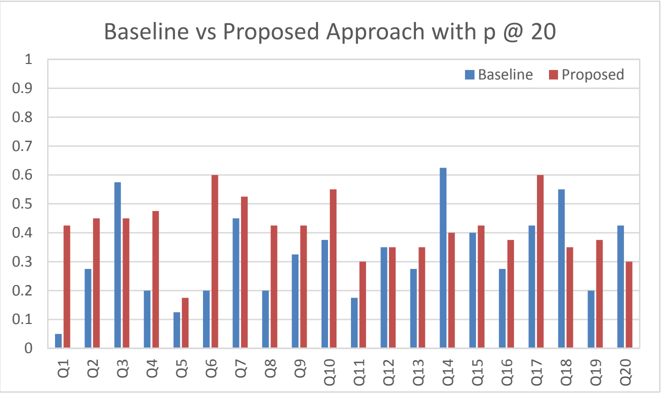

Fig. 4. *p*@ 20 comparison: Baseline vs Proposed retrieval approach

Figure. 3 shows the text retrieval performance of the baseline and the proposed approaches with “Precision @ top 5” (p@5) scores We analyse the top 10 texts retrieved for each query in this section. The retrieval performance has increased for queries Q6, Q11, Q15, Q17 and Q19 significantly. The effects of query drift has been observed for two queries Q13 and Q14. The decrease in the retrieval performance for query 13 is the same as the one described in the above paragraph. For query 14, texts with noisy data degraded the performance with p@10. The overall retrieval task is much improved for the rest of the queries.

Figure. 4 shows the text retrieval performance of the baseline and the proposed approaches with "Precision @ top 5" (p@5) scores. The top 20 text retrieval performance is effective for queries Q1, Q4, Q6, Q8, Q10, Q17 and slightly degraded performance has been observed for a few queries, namely Q3, Q18 and Q20. For query Q14, the texts retrieved at the later part of the top 20 consists of partial matching contexts of the query. So the overall p@20 score reached 0.4 which is better than the retrieval performance of top 10 texts. For query Q17, we observed equal performance with baseline and the proposed method. The proposed method for this query fetched texts having,"54% consumers found mobile payments as quick and easy way" and "majority of customers realized convenience and gained benefits via loyalty programs". In the mean time, the baseline approach fetched texts having similar observations like "customers benefit through the ability to control/monitor their finance" and "36% customers reported mobile payments as the easiest way". However, the retrieval performance is consistent with Q17 whose p@d scores remained on an average of 0.61. Finally we have presented the overall performance of the proposed approach vs the baseline approach with the averaged p@d scores across all 20 queries in Figure 5. Subsequently, we plan to incorporate the reasoning and learning processes to identify relation axioms, as in [15], from the web corpus.

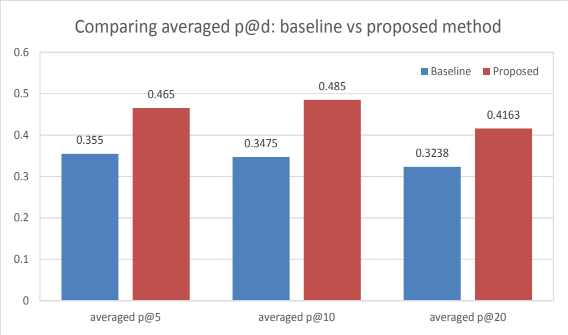

Fig. 5. Averaged p@d comparison: Baseline vs Proposed retrieval approach

7 conclusion

We proposed a framework towards developing an intelligent business information system with multi-faceted data analysis and decision support. As part of the proposed framework, we present an approach for re-ranking of texts with the help of latent topics identified from the web corpus. The proposed approach first performs weighting of query terms and retrieves the initial set of texts from web corpus. We used Latent Dirichlet Allocation method on the same web corpus to find out the topics distribution that cover the contextual features of various financial services/products. The retrieved texts are fused with topic distribution information and re-ranked based on the contextual features. Our experimental results show that the proposed approach captures a variety of key contexts of user information needs in a better way with the support of topics distribution.

Acknowledgments: A part of this work is supported by B-MIDEA project funded by Enterprise Ireland.

References

1. Teece, D.J.: Business models, business strategy and innovation. Long Range Planning **43**(2 - 3) (2010) 172 – 194
2. Horrocks, I.: Ontologies and the semantic web. Commun. ACM **51**(12) (2008) 58–67
3. Osterwalder, A., Pigneur, Y.: Business Model Generation: A Handbook for Visionaries, Game Changers, and Challengers. John Wiley & Sons, USA (2010)
4. Chang, S.K., Jungert, E., Li, X.: A progressive query language and interactive reasoner for information fusion support. Inf. Fusion **8**(1) (jan 2007) 70–83
5. Park, Y.J., Kim, B.C.: An interactive case-based reasoning method considering proximity from the cut-off point. Expert Syst. Appl. **33**(4) (2007) 903–915
6. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3** (2003) 993–1022
7. Blei, D.M.: Probabilistic topic models. Commun. ACM **55**(4) (2012) 77–84
8. Griffiths, T.: Gibbs sampling in the generative model of latent dirichlet allocation. Technical report, Stanford University (2002)
9. Yi, X., Allan, J.: A comparative study of utilizing topic models for information retrieval. In: Proc. of the 31th European Conference on Information Retrieval. ECIR '09, Springer-Verlag (2009) 29–41
10. Bai, J., Nie, J.Y.: Adapting information retrieval to query contexts. Inf. Process. Manage. **44**(6) (November 2008) 1901–1922
11. Liu, X., Croft, W.B.: Cluster-based retrieval using language models. In: Proc. of the 27th ACM SIGIR conf. on Information Retrieval. SIGIR '04, New York, NY, USA, ACM (2004) 186–193
12. Aamodt, A., Plaza, E.: Case-based reasoning: Foundational issues, methodological variations, and system approaches. AI Commun. **7**(1) (1994) 39–59
13. Öztürk, P., Aamodt, A.: A context model for knowledge-intensive case-based reasoning. Int. J. Hum.-Comput. Stud. **48**(3) (1998) 331–355
14. Salton, G., Wong, A., Yang, A.C.S.: A vector space model for automatic indexing. Communications of the ACM **18** (1975) 229–237
15. Sánchez, D., Moreno, A., Del Vasto-Terrientes, L.: Learning relation axioms from text: An automatic web-based approach. Expert Syst. Appl. **39**(5) (2012) 5792–5805