Basora Enterprise Search: A New System for Enterprise Information Retrieval

Maurice G. Wallé and Leon J.M. Rothkrantz

Abstract: *There are several issues at play within Enterprise Search (ES), a major one is that enterprises are finding it difficult to retrieve their data using current ES solutions. This research is a step towards improving this. By designing, implementing and evaluating a system for Enterprise Information Retrieval, with an improved search result presentation technique, which assists users during search tasks. This was accomplished by researching current ES systems and search result presentation techniques. This research resulted in finding two forms of document summaries, namely a textual document summary technique: Top Ranking Sentences (TRS) and a visual document summary technique: Thumbnails. These document summaries have been designed to support a user during information seeking activity. The system developed was named Basora Enterprise Search (BES) and was developed using an agile software development approach. It incorporates the TRS and Thumbnail into its search result presentation technique. The BES prototype was put through both a performance and user evaluation. The first test indicated that BES performs more or less equal to the Commercial Enterprise Search Solution IBM Omnifind 8.5, while providing the user with two extra forms of document summaries. The user evaluation focussed on evaluation the effectiveness of the new search result presentation technique. The results of the user evaluation show that there are various search tasks where the addition of these summary elements has a positive effect on relevance assessment and query reformulation. This research indicates that the BES system actually helps a user assess the relevance of a document, minimizing the amount of documents that need to be opened before the user finds the desired one. It also indicates that the additional visual and textual document summaries assist the user when reformulating a search query, decreasing the time it takes a user to complete a search task.*

Key words: *Enterprise Search, Search Result Visualization, Top Ranking Sentences, TRS, Thumbnails*

INTRODUCTION

Information Retrieval can carried out in several environments, two of these are: Web and Enterprise. For each of these environments a different type of Search Engine is required namely Web Search Engine or Enterprise Search Engine. Web Search entails, using a search engine to identify and retrieve information housed on the World Wide Web. While an Enterprise Search focuses on retrieval of information within Enterprises (i.e. Company's), this research focuses on the Enterprise Search environment. A term within Enterprise Search that is getting a lot of attention is "Findability". Its definition according to Frappaolo & Keldsen 2008 [4] is "The Art and Science of making Content Findable". The reason for this attention is that the volume of content within businesses is growing at a phenomenal pace, and finding relevant content using simple queries is not sufficient, so the need arose for Enterprise search to become a science hence the term "Findability". They also state that the ineffectiveness of Enterprise Findability is not the fault of a poor search engine but the design behind its deployment is flawed. These are some current issues within Enterprise Search, but when looking at this domain from a different angle it is possible to identify several ontology's namely the user, the system and the network. Each of these has its own set of hurdles, within the first ontology the user; there are cognitive aspects like, how a user defines a query, how does a user assess the relevance of a document etc. Furthermore the user interface plays an important role in the last aspect mentioned.

The system has its own hurdles, such as how to convert a user's query into application language so the system understands what to retrieve based on the users query. Another issue is how to represent and store the data for the search engine to access this quickly, keeping the response time to a minimal. The last ontology to consider is the network, one of the reasons is that network lag can seriously affect the response time of the Search Engine. It is also important to consider on what kind of system the search engine runs, because when running it from a PDA the network

response time is slower than when using a computer. The topic of this research is to design, implement and evaluate a system for Enterprise Information Retrieval, which employs an improved search result presentation technique. This research focuses on the user and system aspect of Enterprise Search.

There are three challenges that lie at the centre of this research: The first one is that formulating and re-formulating good search queries is proved to be a cognitively challenging task for users. These search queries are often approximations of a user's underlying need, making it difficult to formulate a query in one try, to satisfy this need; consequently making the information seeking process an iterative one [5, 10, 11]. The second one is interpreting and assessing the relevance of documents in search result list, this is also imperative to the search process [5, 12]. Users are reluctant to examine large numbers of individual documents and they hardly look further than the first result page [5, 13]. So a user's decision to view a document or not is totally dependent on the available elements within the search results namely: document name, author, file type, snippets, etc. Lastly building a system that allows a user to search trough a dataset, which assists with re-formulating search queries and interpreting and assessing the relevance of documents retrieved. The first two challenges are user related while the last is associated with the system ontology.

RELATED WORK

By exploring current Enterprise Search systems, information could be gathered on how to tackle the last research challenge mentioned above. The anatomy of Google's1 architecture was reviewed to get a better understanding of the components of large scale search engines. Google was chosen because their success and ingenuity, which could be used in the development of the new Enterprise Search system. Also the differences between the components of an Enterprise and Web search engine are minimal. These components are shown in Figure 1 and discussed in the following paragraph.

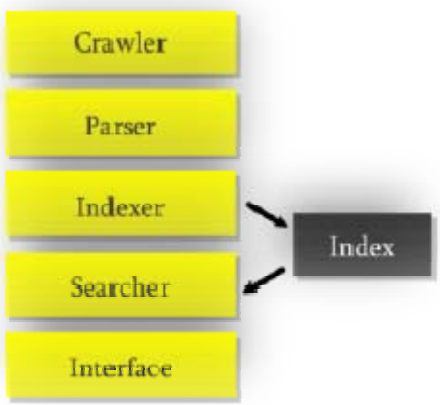

Fig.1. Basic Components of a Search Engine.

Next research was also performed on search result presentation techniques. This research led to finding two forms of document summaries, namely a textual document summary technique: Top Ranking Sentences (TRS) and a visual document summary technique: Thumbnails. TRS are up to three sentences presented in the search results that contain the query terms and are ranked according to several features. Their main purpose is to offer the user extra information about a document, so beforehand they can form a better opinion of what information a document contains [6]. The thumbnail is a miniature image of the first page of a document; this also increases ability of a user to make more accurate decisions about the relevance of search results [7]. These document summaries have been designed to support a user during information seeking activity. These user interface components were chosen for the development of the new

Enterprise Search system, because of the positive results of several studies using one of these or both of these components (Joho and Jose 2006[5], White et al 2005 [6], Tombros and Sanderson 1998 [15] and Dziadosz and Chandrasekar 2002 [7]).

Finally Solr 1.2 Search server2 and IBM Omnifind Search Solution3 were explored for their Indexer and Searcher components. Seeing that the main focus of this research was not to implement these components, and since an implementation already exists and which is well tested, the development of the Indexer and Searcher seemed unnecessary. Solr and IBM Omnifind were researched to be used for these components of the new Enterprise Search system. The Solr Search server was chosen because it's an open source project and could be manipulated to provide all the functionality needed for the prototype; while the IBM Omnifind Solution was less adjustable to the Basora prototype development requirements. It was also very difficult to access the Indexer and Searcher components of the IBM Omnifind Solution for integration into the Basora prototype. Another key factor was that the Solr server, was well documented and tested; the documentation was easily accessible, unlike the documentation for the IBM Omnifind Solution. This was very hard to obtain, often it was incomplete or to superficial. It was possible to contact the IBM support line, but this was usually quite a lengthy process. In the end Solr 1.2 was chosen because of its speed, agility and good documentation.

GLOBAL DESIGN

The design of the BES prototype can be split up into the components seen in Figure 2. Each of these components performs a vital role in order to allow the prototype to function. The Crawling, Parsing and Committing components are also referred to as the Basora Pre-processing stage, because this has to be done before-hand for the prototype to function.

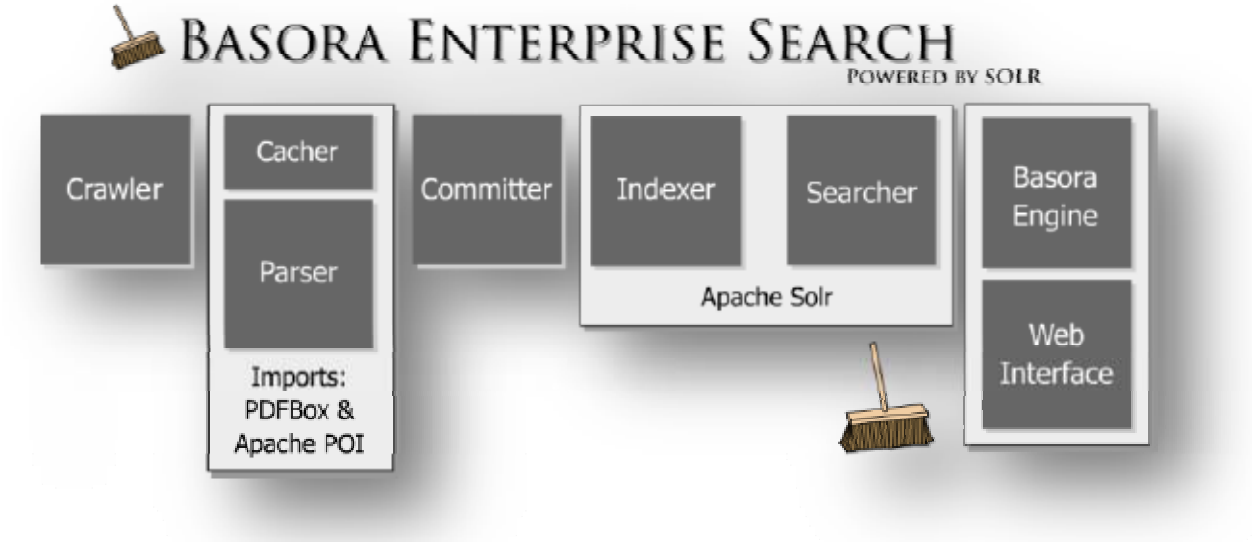

Fig.2. Basora Enterprise Search prototype components.

The Crawler component of the BES system is able to crawl a file system and retrieve the location of all the, Microsoft Word (.DOC), PowerPoint (.PPT), Excel (.XLS) and PDF documents (.PDF), along with the date and time it was last modified and it also assigns each document a unique identifier. This information is used when the file system is re-crawled, so it can check if the file has changed in between crawling sessions. Now the Parser component is responsible for extracting the metadata and text from the crawled documents. This is then formatted this into XML, so it can be committed to the Solr Indexer, because the Indexer only supports XML files. The Committer is the link between the Basora Parser and the Indexer; it is responsible for

sending the Parsed XML files to the Indexer. It reads the parsed XML files one by one and sends them to the Indexer. Which has to index these by storing and caching them; for when a search query is submitted to the Searcher it can quickly retrieve the relevant documents. The Indexer and Searcher components of the BES prototype have been configured using the Solr Search server. These components were used because developing these is outside of the scope of this research and that it is already available and that Solr is a well respected and tested Search Server. The results are returned by the Solr Searcher in XML format, this contains all the necessary data to be able to extract the TRS. This is part of the responsibility of the Basora Engine; its main purpose is to function as a bridge between the Solr Searcher and the Basora Web Interface. A submitted query first goes through the Basora Engine where it is formatted and sent to the Solr Searcher. Then the results it obtains are parsed and formatted so these can be displayed to the user through the Web Interface. This interface offers an improvement over existing Enterprise Search systems. It addresses the first two research challenges (discussed in the first chapter) by adding a visual and a textual document summary to the result presentation namely Top Ranking Sentences (TRS) and a Thumbnail. TRS are up to three sentences presented in the search results that contain the query terms and are ranked according to several features. Their main advantage is offering the user extra information about a document so beforehand they can form a better opinion of a document's contents [6]. The thumbnail is a miniature image of the first page of a document, this increases ability of a user to make more accurate decisions about the relevance of results [7]. These elements are further discussed in the next chapter.

IMPLEMENTATION

The Basora Enterprise Search (BES) prototype's main purpose is to make a dataset searchable while assisting a user with query reformulation and relevance assessment. The design consisted of various components as discussed in the previous chapter. These were implemented using an agile software development approach, incrementally developing BES prototype. The Crawling Parsing and Committer components are the pre-processing steps of the BES system. These were developed from scratch using some essential libraries namely Apache POI & PDFBox (for reading PDF and Office documents). The pre-processing components are responsible for retrieving all documents from a file system, extracting the metadata and content from these, formatting this into indexable XML and finally committing this to the Indexer.

The Apache Solr Server was customized to provide Indexer and Searcher support for the BES system; Solr is a subproject of the Apache Lucene open source Java search library. It ensures that the pre-processed data is properly stored and also once a query is submitted to the Searcher it retrieves the results as quickly as possible. The Basora Engine's main purpose is to allow the Web Interface and the Solr Searcher to communicate. It is also responsible for extracting the TRS, and formatting the results in HTML so they can be displayed to the user through the Web Interface. This user interface allows users to interact with the BES system during the extent of their search. A screenshot of the main interface is shown in Figure 3.

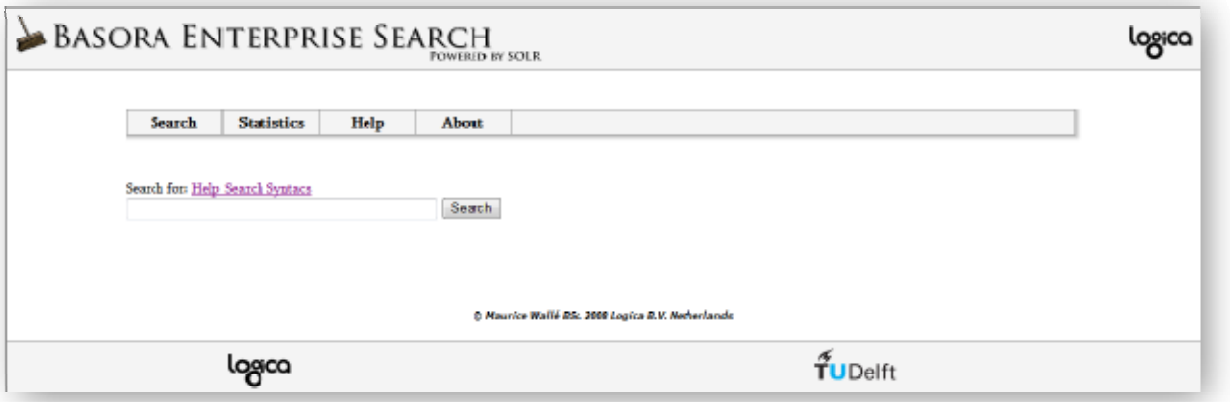

Fig.3. Basora Enterprise Search Main Interface.

TRS ALGORITHM

Seeing that no actual TRS algorithm could be found, this was custom made, inspired by the literature available. The variables were chosen based on the work described in the paper White et al [6]. The system extracts all sentences in which the query terms occur and ranks all these sentences accordingly. The three characteristics used are discussed next.

The first one, "position of the sentence within the document" is determined based on its location within the document. A sentence that occurs in the upper or lower 20% of the document is usually part of the table of contents, the introduction, the executive summary, the abstract, or the conclusion. These components are very important so these sentences get a high ranking. If the sentence is located elsewhere in the document it will be given a low ranking. The proportions were obtained by gathering about 20 documents from the available dataset4 and calculating these. Then the second one, "term occurrence" is calculated by dividing the total occurrence of query terms within all sentences by the total occurrence of all these query terms within each sentence. Then multiplying this by 100 and rounded to an Integer. The final characteristic, "amount of terms per sentence" ranking is only calculated if the user submits two or more query terms to the engine. If this requirement holds, the engine calculates the amount of query terms within a sentence, and divides this by the total amount of query terms supplied by the user. Then multiplies this result by 100 and rounds this to an Integer.

Once these individual rankings have been gathered, the engine calculates the total ranking per sentence based on Equation 1. The equation treats the Position Ranking as the variable with the most influence over the total ranking and the Amount of Terms per Sentence Ranking as the variable with the least. This approach is taken because this characteristic is described in the literature as quite an influential one [6]. The exact values for the total ranking calculation were obtained by iteratively tweaking these. Once the entire algorithm was implemented it was tested to evaluate whether the ranking function actually ranked more relevant sentences higher than less relevant ones. This was done by first choosing five fixed query terms and analysing the sentences and respective ranking obtained. The final values were obtained by tweaking Total Ranking Calculation values (5, 3 & 2) and the Sentence Position variable until the desired results were obtained. This TRS algorithm was implemented in Java 1.6 using the standard Java libraries.

Equation 1. Total Ranking Calculation.

$\frac{(5x + 3y + 2z)}{10}$	x = Position Ranking y = Term Occurrence Ranking z = Amount of terms Ranking

Search Result Presentation Layouts

The BES prototype contains 4 different layouts (illustrated in Figure 4 to Figure 7) for search result presentation. The first contains the same amount of information as the current Enterprise Search Solutions provide to their users, this one is referred to as the baseline. While the other three layouts consist of the baseline components with the addition of either or both, the visual and/or textual document summaries (TRS and Thumbnail). The BES system's Web Interface was developed using the Java 1.6 programming language, Java EE, HTML 4.01, and CSS 2.0.

Fig.4. Layout 1: Baseline.

Fig.5. Layout 2: Baseline & TRS.

Fig.6. Layout 3: Baseline & Thumbnail.

Fig.7. Layout 4: Baseline, TRS & Thumbnail.

PERFORMANCE EVALUATION AND RESULTS

The performance and user evaluation was carried out on the same computer and using the same fixed dataset. The computer's specifications were: Intel Core 2 Duo T7500 2.20 GHz (2 CPU's), 800 MHz FSB, 2 GB of RAM, 100 GB Hard disk drive, and runs Windows XP Professional Edition SP2. The dataset consisted of 5408 documents: 50% Microsoft Word, 5% PowerPoint, 8% Excel and 37% PDF documents. For the performance evaluation the following variables were measured: the time it takes to pre-process the dataset, response time of the search engine and the average number of results it returns. These were taken from an Enterprise Search Engine evaluation guide by Google [14]. This document describes seven critical characteristics of an Enterprise Search Engine. Only three aspects were evaluated, because these were needed to indicate how the Basora Enterprise Search (BES) prototype performs compared to current Enterprise Search Solutions. Some of the other characteristics mentioned in the evaluation guide were: Security, How often it crawls the content, Costs and maintenance fees, these were outside of the scope of this evaluation. These variables were measured for both the BES system using layout 4 and the IBM Omnifind 8.5 Enterprise Search solution (used as benchmark).

Table 1. Pre-processing performance results

Search Engine	Crawling	Parsing	Indexing	Total Time
Basora	4 min	3 hour 42 min	10 min	3 hours 56 min
Basora *No Thumbnail generation*	4 min	1hour 41 min	10 min	1 hour 55 min
Omnifind	35 min	1 hour 40 min	1 hour 30 min	3 hours 45 min

First both systems were installed on the test computer, then both systems were put trough their pre-processing steps and clocked. In Table 1 the pre-processing times are shown, these indicate that the BES and Omnifind systems pre-processing time is more or less equal. There is a big difference between both systems: Crawling and Indexing times, because the exact procedure used within the Omnifind solution is not available one could only speculate why. One of the reasons could be that BES does not contain all the security and back-up features that the commercial solutions have. Another valid reason could be that the BES solution's parsing utility directly generates indexable XML that can be posted to the indexer. This procedure is probably different within the commercial Enterprise Search solution.

Table 2 shows the average response time and average number of results obtained from the search engine. This was carried out on both Enterprise Search systems, by first gathering 30 search queries from experts that continuously query the available dataset for information; then executing these on both the systems and gathering the data. BES's average response time was only 75 milliseconds longer than IBM Omnifind, while providing the user with two extra forms of document summaries. Collins 1994 [6] states that a user considers a process to be real time, if it takes less than 100 milliseconds; so this difference is not noticeable by the users, so they will not get frustrated by the time it takes for the results to be returned. Additionally the BES system returns about 11 times more search results then the IBM Omnifind system; this could be because the Omnifind engine has a stricter filtering function, but it could also indicate that the Omnifind engine indexes less of the documents contents than the Basora prototype does.

Table 2. Average Response Time and Average Number of results returned

Search Engine	Response Time (ms)	Num results returned
Basora	[Solr: 161, Basora Engine: 72] Total: 233	1558
Omnifind	158	133

USER EVALUATION AND RESULTS

The user test was done to evaluate the effectiveness of the search result presentation layouts, in order to assess if the two first research challenges (presented in chapter 1) were met. A total of thirty-two users (6 female and 26 male) participated. This was set up using the IMPACT [1] model for user evaluation. Each participant had to carry out four tasks; each task was done using a different layout (presented in paragraph 4.2). Every task either consisted of a Background search task, Decision making task, Known item task or Topic distillation task. These were formulated based on a simulated work task approach described in Borlund 2000 [9]. To reduce the bias from participants performing the same tasks with the different layouts all in the same order a Graeco-Latin-Square5 arrangement is used. The participants had to fill in an entry, a session and an exit questionnaire.

The first one established the users' age, gender, occupation, Enterprise Search experience and Web Search experience. From this could be gathered that the ages varied from 21 to 50 with an average of 26.9. Their experience with web search engines varied from 5 to 12 years with an average of 9.1 and their experience with Enterprise Search varied from 0 to 10 years with an average of 2.5. From these participants four were experts in the field of Enterprise search, with an average of 5 years of experience within this area, and an average of 10 years of experience using web search. The second one was suppose to gather the data needed to evaluate how easy it was to assess the relevance of a document based on the information available in the layout. This could be used to assess the relevance assessment and also to gather how much each of the layout elements6 contributed to the user opening a certain document. Finally the last questionnaire allowed the users to vote for their preferred layout and this also gave them an opportunity to comment on the evaluation and also the result presentation layouts.

In order to assess if the new system helps a user during query reformulation and relevance assessment the following measures were used: User interaction, Relevance assessment, Contribution of layout elements and Layout preference. These were used to tackle the first two research challenges (presented in chapter 1)) each of them is discussed in the following paragraphs.

User Interaction

Table 3 consists of 7 columns, the first two indicate the layout that was used and the average amount of queries that were submitted by the participants overall. The third and fourth columns represent the average query length; and the total amount of result pages that were viewed. Next the fifth and sixth columns presents the amount of result pages that were viewed per search query that was submitted and the amount of documents that were clicked per result page that was viewed. Finally the column shows the average amount of time it took the participants to complete a task using a certain layout. Each of these measures are averages taken over the amount of participants namely 32 (n = 32), except for the totals because the n here is 128. The values in each of the cells are the average and the standard deviation (between brackets). The original

values were obtained during the evaluation procedure using a custom BES evaluation add-on.

Table 3. User Interaction

Layout	Queries submitted	Query length	Result pages viewed	Result pages per Query	Clicks per Result page	Time (min)
1	4.72 (2.32)	3.02 (0.94)	6.06 (3.60)	1.28 (0.39)	0.70 (0.47)	5.48 (2.41)
2	5.13 (2.69)	2.82 (0.81)	6.81 (4.21)	1.34 (0.48)	0.58 (0.41)	5.53 (2.16)
3	4.78 (2.43)	3.03 (1.06)	5.81 (3.69)	1.20 (0.29)	0.35 (0.32)	4.97 (2.27)
4	4.25 (2.82)	3.14 (1.85)	5.38 (3.87)	1.51 (1.77)	0.57 (0.60)	5.29 (2.75)
Total	**4.72 (2.56)**	**3.00 (1.22)**	**6.02 (3.84)**	**1.33 (0.94)**	**0.55 (0.47)**	**5.32 (2.39)**
n = 32 (Layout 1 to 4), n = 128 (Total)						

The query length column indicates that longer queries were submitted when using the fourth layout so the users formulated more specific queries when using this layout, this could suggest that they used the extra information available to reformulate more specific queries. The most documents were clicked when using the first layout; this indicates that the information provided in the baseline was not sufficient so more often they had to open a document to find out if this was relevant for their search. This was also seen during the user evaluation that users clicked fewer documents when using the layouts 2-4. Another interesting factor is that seeing that the underlying search engine is identical for all layouts and the users clicked fewer documents per result page when using layouts 2-4. It can be suggested that the participants interacted with the system more often when using the document summary layouts then when using the baseline layout. So TRS and thumbnails appear to increase the interaction between the user and search engine. From these results nothing concrete can be concluded because the standard deviations are all quite large, meaning that the results are spread broadly over the result space. Also the differences between the average values between the layouts are not very big, these only suggest slight differences. This may also be because of the assignments used; it is very difficult to set up assignments, which actually give a good representation of reality.

Relevance Assessment

This paragraph presents the participants relevance assessment from three perspectives as shown in the table below from the left to the right column respectively: "Ease of finding information using the current layout", "Easy of finding new Information with the current layout after query reformulation" and "Easy with which the user could predict the contents of the documents based on the information available the layout". For each of the perspectives the participants were asked to fill in a score, between 1 and 7, 1 indicating best/easiest and 7 worst/hardest. The values in each of the cells are the average and the standard deviation (between brackets).

Table 4. Relevance Assessment

Layout	Ease of Finding	New Information	Contents Prediction
1	2.72 (1.28)	3.28 (1.08)	2.41 (1.48)
2	2.72 (1.35)	2.88 (1.16)	2.28 (1.20)
3	1.97 (1.26)	2.69 (1.33)	2.03 (1.18)
4	1.78 (0.79)	2.59 (1.16)	1.78 (1.01)
Total	**2.3 (1.20)**	**2.86 (1.20)**	**2.13 (1.24)**
n = 32 (Layout 1 to 4), n = 128 (Total)			

The ease of finding is greater using the third and fourth layout; this suggests that the thumbnail contributed the most in facilitating the ease of finding for the user. While for the baseline and second layout the value is the same, this indicates that the ease of finding with the second layout is equal to the baseline, so the TRS do not contribute to this measure. The layouts 2-4 are preferred over the baseline layout when reformulating a query to attain new information. There is a definite difference between the baseline and the document summary layouts, which indicates that these assist a user with query reformulation. For contents prediction the results show that layout 4 is best followed by layout 3 & 2 and finally layout 1. This indicates that the contents prediction increases as the elements TRS and Thumbnails are added to the interface, this suggests that the users prediction accuracy increases as more information is available, which is quite a logical occurrence. This was also noted during the user evaluation that the users could complete several assignments faster, because the extra information allowed them to form better opinions on what a document contained. For each of the perspectives the best rated overall is it the fourth layout, this suggests that this is the most optimal layout and maximizes the user's relevance assessment.

Contribution of Layout Elements

In this paragraph the results obtained from the user tests concerning the contribution of the layout elements are presented. Each of the layout elements got a score from the participants that corroborated how much each of them contributed to the initial relevance assessment. The scores assigned were between 1 and 7, 1 indicating very much and 7 very little. The following layout elements were evaluated: Filename, Author, Snippet, TRS, Thumbnail, File type and File size. The results are shown in Table 5 where a strong contribution is represented by a low score. The values in each of the cells are the average and the standard deviation (between brackets). Note that the sample sizes differ across the layout features seeing that the layout elements TRS and Thumbnail were not available in all the layouts.

Table 5. Contribution of layout elements

Task	Filename	Author	Snippet	TRS	Thumbnail	File type	File size
1	2.34 (1.64)	3.38 (1.88)	2.66 (1.21)	3.88 (1.63)	4.06 (1.61)	6.19 (1.40)	6.66 (1.10)
2	3.13 (1.72)	3.75 (2.08)	2.84 (1.92)	3.00 (1.41)	5.31 (1.99)	6.22 (1.39)	6.59 (1.13)
3	2.06 (1.72)	2.28 (1.65)	4.00 (2.00)	4.56 (1.46)	1.25 (0.45)	3.06 (2.05)	3.31 (2.35)
4	2.78 (2.01)	3.00 (1.97)	4.34 (1.86)	3.50 (1.41)	1.00 (0.00)	2.13 (1.56)	3.69 (2.46)
Total	**2.58 (1.86)**	**3.1 (1.96)**	**3.46 (1.90)**	**3.73 (1.56)**	**2.91 (2.24)**	**4.40 (2.44)**	**5.06 (2.43)**
n = 16 (for TRS and Thumbnail in Task 1 to 4), n = 32 (the rest in Task 1 to 4), n = 64 (for TRS and Thumbnail in Total), n = 128 (the rest in Total)							

From the total row can be concluded that the participants found the filename was the strongest factor when determining whether or not to open a document. While the thumbnail comes second, again demonstrating its importance as in the previous paragraph. The difference between the TRS and the Snippet is not very significant but it does indicate that users tend to use the Snippet more. This could be because the participants are accustomed to the Snippet being a crucial element within the result presentation, and that they still have to get use to the TRS. This was also mentioned by several participants after the evaluation.

When the TRS and Thumbnail results are compared it is clear that TRS is given a stronger score for task 2 a decision making task, while the thumbnail is given a stronger score in task 4 a known item task. This suggests that the effectiveness of TRS and Thumbnails varies across tasks. And this also indicates that TRS may be more useful where Thumbnails are less effective, and vice versa. Another interesting observation is that the Thumbnail is given a very high score for the tasks 3 and 4, while the TRS is given its strongest score for task 2. This can be explained because for task 3 and 4 the user had to find documents of which the front page was known, so the thumbnail made finding the documents a breeze. While for task two the user had to search trough the dataset to figure out whether or not a project had been carried out or not, so for this task having extra information of the TRS was very useful.

A very interesting observation is that the author element was not rated very high for the first task where the objective was to find the name of the author corresponding to a certain project and vice versa. This was mainly because the author field information is extracted from the metadata that is gathered from the document and this is not always completely accurate. Sometimes it contains nothing or a useless variable. It could be suggested that a better means of acquiring the information stored in the metadata should be considered, or the use of metadata should be encouraged or facilitated.

Layout Preference and Participant Comments

Once the participants had completed all the tasks they were asked to rank all the layouts according to their preference. The scores 1 to 4 were used, 1 being the one preferred most and 4 least. In the figure and table below the results are presented. The clear winner among the layouts is layout four with an average of 1.5 followed by layout 3 as the runner up. Twenty of the thirty-two participants choose layout 4 as the preferred layout, while twenty-three participants choose layout 1 as the least preferred one. Some of the users did not rank layout 4 as the winner because they think it is a bit too cluttered, and they are used to the current search result presentation. This suggests that the new layouts take some getting used to.

Table 6. Layout preference

Preference	Layout 1	Layout 2	Layout 3	Layout 4
1 (Most)	0	1	11	20
2	6	4	13	9
3	3	20	7	2
4 (Least)	23	7	1	1
Average rank	3.53	3.03	1.94	1.50

An interesting observation was that layout preference is dependent on the type of search task that needs to be completed. If a user is looking for a PowerPoint presentation and they remember the image on the front page of the presentation it is very handy if there is a thumbnail available. One of the most noticeable comments was

that TRS were experienced as being very handy, but the users did have to get use to the idea of them. Some users also thought this was a redundant element, others thought that maybe the TRS should replace the snippet. The thumbnails seemed to be the participants preferred addition; this indicates that this element can be very useful when performing a particular type of search. One of the experts implied that is was very obvious how for each of the evaluation tasks an optimal layout exists, showing that maybe a user should be able to choose a layout based on the search task that has to be carried out. The overall positive performance of layout 4 could also be that it offers a support in a wider range of tasks that layout 2 or 3.

CONCLUSIONS

This research demonstrates that this new Enterprise Document Retrieval system, Basora Enterprise Search actually helps a user assess the relevance of a document easier minimizing the amount of documents that need to be opened before the user finds the desired one. It also suggests that the additional visual and textual document summaries assist the user when reformulating a search query, also decreasing the time it will take a user to complete a search task. And finally it also indicates that it is possible to build a system that performs more or less equal to the to the current response time standards of Enterprise Search Engines, containing these extra useful components.

Implications

The results from this research project have several implications for the design of Enterprise Search Engine interfaces. Especially because they demonstrate that the addition of these textual and visual document summary elements has a positive effect on relevance assessment and query reformulation. Also because the system performs more or less equal to the to the current response time standards of Enterprise Search Engines, makes is attractive to consider adding this improved search result visualization technique to current Enterprise Search Interfaces. The participants of the user evaluation often found it easier to find relevant documents and new information when TRS and Thumbnail elements were added to the search result interface so this suggests that the current Enterprise Search engine's result presentation is not necessarily optimised and that there is room for improvement. The BES system is not completely optimized, especially the pre-processing steps parsing component because this relies on open source Java libraries that are work in progress. Another minor issue is that support should also be built in for the new Office document format, seeing that its popularity is increasing. These are some of the aspects that still need optimization but even so the system demonstrates that it can improve a user's search experience. It is not the intention of this research project to suggest that the current interfaces are useless, because this is surely not the case seeing that these have been used years and they are effective. But the intention is to show that the evolutionary cycle of Enterprise Search Engine Interface design, has reached a new stage. Serious considerations should be made in favour of this advance, because the addition of these document summaries not only facilitates query reformulation and relevance assessment. But also increases the level of interaction between the user and the search engine interface; also assisting in closing a gap the science of Human Computer Interaction.

REFERENCES

[1] Benyon, D. and Turner, P. and Turner, S. 2005, "Designing Interactive Systems", 1st edition, Harlow Essex England. 2005.

[2] Langville, A.N. and Meyer, C.D. 2006, "Google's PageRank and Beyond: The Science of, 1st edition, Princeton University Press, Woodstock Oxfordshire England, 2006

[3] Collins, D. 1994, "Designing Object-Oriented User Interfaces", 1st edition, The Benjamin/Cummings Publishing Company Inc. Redwood City, California, 1994

Papers

[4] Frappaolo, C. and Keldsen, D. 2008, "Findability: The Art and Science of Making Content Easy to Find", in Market IQ Intelligence Quarterly Q2 2008, AIIM, 2008.

[5] Joho, H. and Jose, J.M. 2006. "A Comparative Study of the Effectiveness of Search Result Presentation on the Web", In: Proceedings of the 28th European Conference on Information Retrieval, volume 3936 of Lecture Notes in Computer Science, London, UK, April 10-12, 2006, pages 302-313, Springer-Verslag, 2006.

[6] White, R.W. and Jose, J.M. and Ruthven, I. 2005, "Using Top-Ranking Sentences to Facilitate Effective Information Access", Journal of the American Society for Information Science and Technology, 2005.

[7] Dziadosz, S. and Chandrasekar, R. 2002, "Do Thumbnail Previews Help Users Make Better Relevance Decisions About Web Search Results?", in proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval , 2002.

[8] Brin, S. and Page, L. 1998. "The anatomy of a large-scale hypertextual Web search engine", in journal of Computer Networks and ISDN Systems, Elsevier, 1998

[9] Borlund, P. 2000. "Experimental components for the evaluation of interactive information retrieval systems", In: Journal of Documentation, MCB UP Ltd, Volume 56, Issue 1, pages 71-90. 2000.

[10] Belkin, N.J. and Oddy, R.N. and Brooks, H.M. 1982, "ASK for information retrieval: Part I. Background and theory". Journal of Documentation, 38(2): pages 61-71, 1982.

[11] Belew, R. 2000, "Finding Out About – Search Engine Technology from a Cognitive Perspective", Cambridge University Press, 2000.

[12] Hearst, M.A. and Pederson, J.O. 1996, "Re-examining the Cluster Hypothesis: Scatter/Gather on Retrieval Results", in Proceedings of the 19th annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Zurich, Switzerland: ACM 1996.

[13] Jansen, B.J. et al 1998, "Real Life Information Retrieval: A Study of User Queries on the Web", ACM SIGIR Forum: A Publication of the Special Interest Group on Information Retrieval, 32(1): pages 5-17, 1998.

[14] Enterprise Search Evaluation Guide, "Seven critical characteristics your enterprise search solution must have", 2006.

[15] Tombros, A. and Ruthven, I. 1998, "Advantages of query-biased summaries in information retrieval" , in Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Melbourne, Australia: ACM. 1998.

ABOUT THE AUTHORS

Prof. Leon Rothkrantz, drs.dr., Faculty of Electrical Engineering, Mathematics and Computer Science, Man Machine Interaction Group, Delft University of Technology, Phone: +332787504, E-Mail: L.J.M.Rothkrantz@tudelft.nl

Netherlands Defence Academy, Faculty Military Sciences, Den Helder, The Netherlands.

Maurice G. Wallé, MSc, Faculty of Electrical Engineering, Mathematics and Computer Science, Man Machine Interaction Group, Delft University of Technology, Phone: +332787504, E-Mail: M.G.WALLE@student.tudelft.nl