Experiences with Automated Categorization in E-Government Information Retrieval

Tanja Svarre* and Marianne Lykke**

*/** Dept. of Communication and Psychology, Aalborg University,
Nyhavnsgade 14, DK-9000 Aalborg, Denmark,
* <tanjasj@hum.aau.dk>, **<mlykke@hum.aau.dk>

Marianne Lykke is Professor and Knowledge Group leader for the e-Learning Lab (eLL), Department of Communication and Psychology, Aalborg University. She is Professor II at the Oslo and Akershus University College of Applied Sciences (HIOA), and is visiting professor at Åbo Academy University, Åbo. Her research concerns technologies for knowledge sharing and learning in organizations, specifically information architecture and interaction design, and use practice. She is member of several editorial boards, and has published in international as well as national journals, anthologies and proceedings. She has acted as consultant to many enterprises and government organizations.

Tanja Svarre is Assistant Professor at the e-Learning Lab, Department of Communication and Psychology, Aalborg University. Her research is centered on professional information practice and design of ICT-based services with a specific focus on information architecture. She defended her thesis at Aalborg University in 2012. Here she investigated and compared indexing methods in e-government from a user based perspective.

Svarre, Tanja and Lykke, Marianne. **Experiences with Automated Categorization in E-Government Information Retrieval.** *Knowledge Organization. 41(1)*, 76-84. 35 references.

Abstract: High-precision search results are essential for supporting e-government employees' information tasks. Prior studies have shown that existing features of e-government retrieval systems need improvement in terms of search facilities (e.g., Goh et al. 2008), navigation (e.g., de Jong and Lentz 2006) and metadata (e.g., Kopackova, Michalek and Cejna 2010). This paper investigates how automated categorization can enhance information organization and retrieval, and presents the results of a realistic evaluation that compared automated categorization with free text indexing of the government intranet used by Danish tax authorities. The evaluation demonstrates a potential for automated categorization in a government context. In terms of quantitative measures free text indexing performed at the same level or better than searching by categorization. However, the qualitative analysis revealed that categorized overviews were useful if the participant did not possess much knowledge of the task at hand. When task knowledge was present, categorization was used to support the assumptions of a correct search. Participants avoided automated categorization if high-precision documents were among the top results or if few documents were retrieved. The findings emphasise the importance of simultaneous search options for e-government IR systems, and reveal that automated categorization is valuable in improving search facilities in e-government.

Received 31 July 2013; Revised 20 September 2013; Accepted 24 September 2013

Keywords: search, categorization, categories, queries, documents, E-government, information retrieval

1.0 Introduction

E-government facilitates governments utilising ICT to communicate with and allow access to information for stakeholders (e.g., Fang 2002; Jaeger 2003; Grant and Chau 2005). Documentary support is essential for operations undertaken in public administrations (Kraemer and Dedrick 1997; Klischewski 2006; Sabucedo and Rifón 2006). Therefore, not being able to find needed information can have severe human and financial costs (Kraemer and Dedrick 1997). Different tools add to reduced information overload in organizations, e.g. value added information (Edmunds and Morris 2000). Metadata assignment supports interoperability between systems, high precision search, and knowledge sharing (Schwartz, Divitini and Brasethvik 2000; Moen 2001; Choo 2006; Tambouris,

Manouselis and Costopoulou 2007). Metadata can be assigned manually by humans or automatically based on a machine-generated analysis of documents. In Danish e-government, the predominant approach is manual assignment (The Danish Government, Local Government Denmark (LGDK) and Danish Regions 2007).

In the field of US federal records management, Sprehe et al. (2002) found that different situational factors affected the quality of federal employees' record-keeping, causing a divergence in the quality of record management across governments. Factors like availability of resources and guidance, the motivation of employees, and efficiency of access to records appeared to affect the quality of records management in the study. In a recent study of metadata assignment in a Finnish government, the researchers found that employees preferred not to assign metadata when they had the option. Additionally, the employees tended to accept default values whenever they were available (Kettunen and Henttonen 2010). The results suggest that e-government indexing can benefit from an automatic solution to indexing in a number of ways. The literature has already demonstrated that the assignment of metadata is one among a number of prerequisites for retrieval and sharing of knowledge in organizations (e.g., Choo 2006). If automated assignment can improve subject metadata, then we can assume that retrieval and knowledge sharing is also influenced in a positive sense.

2.0 Categorization

Categorization places documents in categories, usually in a web-based environment, with the purpose of supporting searches (Qi and Davison 2009). Specifically, categorization enables post-limitation of search results on the basis of document characteristics, e.g. subject, document type, authors, etc. Categorization may be based on either manually added metadata or automated procedures. Automated procedures include clustering, knowledge engineering and machine learning. Clustering is an unsupervised procedure. Here digitalised documents are represented as document vectors. Calculations of the similarity between vectors subsequently form the basis of clustering documents with corresponding characteristics (Carpineto et al. 2009). Knowledge engineering and machine learning are typically based on a coupling between documents and a controlled vocabulary. Knowledge engineering is a rule-based approach. The rules ensure automated placement of documents in one or more correct categories. The development of rules is done manually. Machine learning on the other hand is based on supervised training. A set of training documents representing each category in the controlled vocabulary is selected and subsequently used for categorization of the full collection of documents (Sebastiani 2002).

Automated categorization has been thoroughly evaluated in individual studies and in comparative reviews. However, the evaluations have to a large extent been system-driven and included no users or had a very limited inclusion of users. Early examples include Apté, Damerau and Weiss (1994), Chen (1995) and Dumais et al. (1998). Turmo et al. (2006), Chung et al. (2010), and Qu et al. (2012) are more recent examples. Zamir and Etzioni's (1999) evaluation of their cluster-based interface Grouper is one example of a user-based evaluation. They found that users explored several clusters to locate relevant documents and that the Grouper users found more documents compared to the baseline system (HuskySearch). Another example is Kules and Shneiderman's (2004) study. They made a comparative study of ranked and categorized outputs in U.S. government webpages. The participants find the overview easy to use and helpful in noticing areas not covered by search results. The authors also note a learning effect from the categorization. Despite the controlled character of the test, the authors conclude that categorization is useful in supporting understanding of large sets of search results.

Lastly, Käki (2005a; 2005b; Käki and Aula 2005) built the evaluation of a web categorization interface (extracted indexing) on users. Different evaluations have been reported from the study. Käki and Aula (2005) made a comparative study of an interface comprising the algorithm and categorized search interface with the World Wide Web as the test base. The study found that the categorized interface had a better average performance in precision (62% against 49%) and recall (33% against 19%). A longitudinal study elaborated on the initial results (Käki 2005b). It was found that categories were used to select 26% of the accessed result pages. The participants indicated that categories were useful, when "the original query was vague, broad, general, or contained words that have multiple meanings" (Käki 2005b, 138). Also, categories helped increasing the focus of a less precise query and were found useful when result rankings were deficient. The results of the study are interesting because they demonstrate that categorization is not necessarily useful in all information searching situations. From the analysis, we get an indication of situations in which categories may be useful. However, a more systematic investigation would be relevant.

Many studies have examined various forms of automated categorization, but few with the participation of users. In the present paper, we investigate automated categorization based on a controlled vocabulary applied with a combination of machine learning and knowledge engineering. We evaluate the automated categorization approach on a corporate and e-government intranet by including professional users. The evaluation is carried out as a comparison study between automated categorization

and automatic free text indexing. On this basis, the research question guiding our further work runs as follows: What characterizes the potential role of categorization in professional e-government information retrieval?

3.0 Methodology

For answering the research question, we carried out a search test in a realistic setting in a real life government intranet at the Danish Tax Corporation, SKAT. The test took place in June 2010 in two office locations of SKAT. The organization intranet contains a heterogeneous collection of documents, e.g. legal directions, citizen and business directions and brochures, legal documents, forms, news, minutes, job postings, reports from finished internal projects, HR information and other internal information from the organization and departments. At the time of the test, the intranet contained 681,640 documents. The search test compares free text indexing, (extracted indexing, system A, baseline) and categorization (assigned indexing, system B) in an experimental manner to be able to observe and capture differences in searching behaviour between the two systems.

A prototype of the organization's future intranet functioned as the test system of the search test. The test system contained a random sample of the running intranet. The sample comprised 188.600 documents, that is, 28% of the full document collection. The prototype was based on content management technology. Autonomy's (www.autonomy.com) search software, IDOL, provided the search functionalities of the search interface. Though more fields were available, the participants only used the search field's query box, search operator and document type during testing. Search results were relevance ranked. For each hit, the document title, a snippet highlighting the search words and the surrounding words, the document type and the date of publication appeared.

System B represented searching by categorization. In IDOL, categorization is based on machine learning. The taxonomy used for the categorization has 169 terms divided into two levels. The selection of one or more categories took place after a search had been processed and a result existed. On the basis of the retrieved documents, the search result was limited to subjects present in the search results. The categorization window displayed the terms from the taxonomy actually containing documents in the current result set. In the test situation, when the participants used system A, the categorization field was covered.

The development of the test database and the training of the document categorization were still taking place during the test work. Consequently, the test work was challenged in various ways. The categorization procedure was semiautomatic, as a part of the documents were placed in the categorization on the basis of manually added subject metadata (documents published after January 1 2008.) The remainder of the documents was indexed automatically. Also, there was a lack of most recent documents. The test database was generated in August 2009 and was not updated in the intervening period of time up to the search test in June 2010. Lastly, the test database had some functional inexpediences, e.g. not being able to link to the full text of all documents and at times slow responses. The test procedure was designed with these inexpediencies in mind to reduce the influence on the test outcome.

Thirty-two employees participated in the test. The participants were recruited by e-mail. In our selection of participants, we emphasised frequent intranet use and information seeking. Forty-two of the voluntary employees met the requirements. Of these 10 were used as pilot testers. We employed three simulated and one genuine work task in the test (cf. Borlund 2003). The simulated tasks covered the sale of an apartment (sim1), taxation of e-commerce (sim2), and tax-based issues related to freelance work (sim3). The test procedure consisted of: 1) an introduction to the session; 2) the search part in which the participants carried out searches in the two systems; and, 3) a post-search interview. In the first part, the participants were introduced to the session, system characteristics, etc. Due to time constraints the participants did not try out the prototype ahead of the test. In all test sessions, the succession of tasks and systems were rotated. When searching in system B, the participants were obliged to use categorization for limiting their search results. The relevance of retrieved documents was assessed on the basis of the title and snippet. The relevance of search results was noted when the result lists appeared. After the search part, a short post-search interview was conducted. The test sessions ranged between 30 minutes and two hours. The test setting (recruitment e-mail, search tasks and the general test session) was pilot tested ahead of data collection.

Different data were collected throughout the search test. The participants' interaction with the test system was logged using the software Morae (http://www.techsmith.com/morae.asp). Interviews, both oral and in questionnaire form, were carried out along during the course of the search test. Documents' relevance was assessed during the test. Relevance was assessed on a four-point scale. 0 represented not relevant; 1 pointing to the subject, but only by a sentence or the like; 2 denoted a document pointing to the topic, but only by parts and not the full document, and; 3 represented a thorough discussion of the question at hand. The scale reflects Sormunen's (2002) four-point scale. A search log registered search time and words applied. From the screen video recorded during the searches, we manually drew the number of hits retrieved, selection of subject categories, use of information filters and search types. All

were registered in SPSS for analysis (http://www-01.ibm.com/software/analytics/spss/) along with the relevance assessments. Subsequently, statistical analyses were carried out, consisting of univariate and bivariate statistics, frequencies, means and correlations. In the analysis, query success designated a query retrieving at least one document with a relevance measure of 2 or above. Session success was the label for a session that contained at least one successful query. We used the log data to compare system A and system B by the number of concepts applied in queries and the degree of search success in sessions and queries. Also the search log provided detailed data on the extent and types of reformulations carried out. However, qualitative data was needed to understand and explain the patterns identified in the search log, as several iterations and changes of search moves can, but not necessarily do equal a bad session. For that purpose we used a Dictaphone to record the search test and the post interview. The recordings were subsequently transcribed. We used atlas.ti for analysing the interview transcripts.

4.0 Results

The search test provides data on the searching behaviour in the two test systems, system A and system B. In total 128 sessions consisting of 564 queries were undertaken by the 32 participants in 64 sessions in each of the two systems. Table 1 summarises the general findings. The average number of concepts is slightly higher in system B (1.90) compared to system A (1.67). This corresponds to the concept (search key) averages of 1.8 and 1.5 found in Lykke et al. (2012). Further, in a study comparing categorized searches with non-categorized searches, Käki (2005b, 136) found an average of 2.10 search words for the former and 2.04 for the latter. Though Käki investigates search words and we report concepts, the respective results agree that on average more search words are applied in categorized queries than in non-categorized queries.

Variables	System A Sessions N=64 Queries N=229	System B Sessions N=64 Queries N=335
Number of concepts in queries (averages)	1.67	1.90
Number of sessions with reformulations (percentages)	65.6	82.8
Number of reformulations in sessions (averages)	2.58	4.23
Query success (percentages)	30.6	21.5
Session success (percentages)	89.1	84.4

Table 1. General Findings of Variables in Search Test

Reformulations took place in both systems. However, in system A the share of sessions with reformulations was 65.6%, while 82.8% of the sessions in system B required reformulations. In addition, the average number of reformulations was notably higher in system B (4.23) compared to system A (2.58). This means that an average session in system A contains 3.58 queries, while the corresponding number for system B is 5.23. The averages are slightly above the findings of similar studies of web search engines and web portals. Lykke et al. (2012) found an average of 2.5 and 3.2 queries per session. Koshman et al.'s (2006, 1879) average was marginally higher at 3.37. To sum up, the present study, and in particular system B, has an increased number of queries in sessions compared to similar studies. We ascribe the increased number of queries in sessions to the participants' lack of experience with the test system. The lack of experience may also explain the increased success rate at query and session level in system A.

4.1 Reformulations

The type of reformulation adds to our understanding of the search actions carried out by the participants. We analysed reformulations to discover if the category, the search words, the document type or the search operator were changed, if several parameters were changed or if no reformulation occurred (see Table 2). In system A, the overall preferred reformulation is a change of search words. This is followed by a change of the document type and simultaneous change of two or more parameters. Compared to system B, the use of the document type filter is far more common in system A, likely because this is the only possible way of reducing search results in system A without changing the search words or the search operator. Thus, the participants actually used the available options for modification of their search results. Furthermore, the regular use of the document type filter emphasises the importance and relevance of the filter. In system B, the preferred reformulation was a change of categories; this was closely followed by a combination of two or more parameters. Next, a change of query words followed. Document type and search operators were rarely used as query modifiers. It is evident that categories are important, which is to be expected, as they were mandatory in system B. In addition, categories were combined with other parameters to a large extent. Most commonly, a change of category was combined with a change of search words. This reflects the design of the system, where only categories with content were shown to the searchers. Thus, when search words were changed, a change of available categories was likely to occur, as the categories reflected the list of retrieved documents. This also explains the importance of a change of query words as a reformulation.

	Total	
	System A	System B
No reformulations	69 (30.1)	62 (18.5)
Category	-	114 (34.0)
Query words	97 (42.4)	47 (14.0)
Document type	28 (12.2)	8 (2.4)
Search operators	8 (3.5)	5 (1.5)
>1 types simultaneously	27 (11.8)	99 (29.6)
Total	229 (100)	335 (100)

Table 2. Types of Reformulations for All Queries (Percentages)

4.2 Combined system B sessions and queries

During the course of the search test, participants occasionally ended up assessing documents before choosing a category in system B queries. This behaviour had different causes. One was the speed of the system. Thus, in the time waiting for the system to categorize search results, some participants began to review the documents found on the basis of the initial query. On other occasions, the participants saw the document they were looking for in the results list before even deciding on a category by which to reduce search results, and they ended up assessing the initial search results without filtering them by category. We denote these searches as combined system B queries, because users had the intention of using system B but then switched to system A. Likewise, 'combined system B sessions' refers to the sessions that should have been carried out in system B, but participants assessed the relevance of documents found in system A and in system B. The following quotation serves as an illustration of combined system B searches:

> But the first time I searched, I got an e-commerce handbook. I would have preferred that to going down there ["down there" refers to the categorization window on the right hand side of the screen] (P10).

In several cases, when a highly relevant document had been discovered before the choice of a category in system B, the participants could not locate the document in the categories, which occasionally led to frustration:

> It is just as bad, because it says "arrears" and "employers", and it is neither of them. So let's see about "employers"… because it says "employers and A-taxes" And it is withhold by the A-taxes, just like our employers withhold our taxes. I simply can't find it. I know it is in there. But on the basis of this, I can't get in there because when I know where it is at, I would go directly for it instead. (P05).

A third type of behaviour also triggered combined system B queries. When the initial query resulted in very few search results, it did not seem natural to the participants to further reduce already limited search results. Some participants undertook the categorization despite the few results, while others omitted the categorization and assessed the results retrieved on the basis of the remaining search possibilities.

> It says just that ... the costs to the European border should be included in the customs value. The other one regarding transportation, I can see that it is explained with great precision. But in this case, I did not search for "customs" down here [in the categories]. I got it by searching for freight and customs value and "pages with all words." And then I got the customs guidance, which is also the one referring to the customs codes treating the rules about the amount of carriage to add. So this [document] is a three then. But I didn't get it by searching for "business imports" or "shipping" or "exports" [referring to categories] (P32).

The quotation illustrates, in a combined system B query with just two retrieval results, how the participant ends up assessing the documents retrieved without categorization. This supports the assumption put forward by Kules and Schneiderman (2004, 2) that search results must have a certain size to make categorization useful.

	Number of sessions in system B	Number of successful sessions system B
System B	26 (40.6)	22 (40.7)
Combined system B sessions	38 (59.4)	32 (59.3)
Total	64 (100.0)	54 (100.0)

Table 3. Sessions Carried Out in System B or in a combination of System B and System A: frequency and success (Percentages)

The combined system B queries and sessions were coded as system B searches inasmuch as the participants had access to the taxonomy and could be influenced by it. However, in respect of the methodology, an overview of the extent of the queries must be provided. To do this, additional codes were added to enable separation from the correct system B queries. The quote illustrates, in a combined system B query with just two retrieval results, how the participant ends up assessing the documents retrieved without categorization. This supports the assumption put forward by Kules and Schneiderman (2004, 2) that search results

must have a certain size to make categorization useful. Table 4 lists the share of combined system B sessions. The table shows that about 60% of system B sessions contained one or more queries that omitted categories. It is evident from the table that approximately 60% of the successful sessions in system B had at least one query that did not include the choice of a category. The sessions that to some degree pass over the categorization are therefore substantial.

Table 4 enlarges on combined system B sessions. The table shows the system delivering successful results for queries contained in sessions. In that way, the table addresses the sessions based on a combination of the two test systems. It is identified that although a combined system B session included queries conducted in system A and system B, both systems have not necessarily provided useful search results. The share of successful sessions is fairly even between the two systems. Thirteen sessions were solved by omitting categories, and 15 sessions had success in including the categories in their queries. Only four sessions found relevant documents by means of both systems. This means that at the session level, the share of success is fairly even between the two systems. It also means that the participants may have omitted the categorization in some queries of a session, but it may still be that relevant documents are found by means of categorization.

	Frequency	Per cent
Task not solved	6	15.8
System A	13	34.2
System B	15	39.5
Both systems applied	4	10.5
Total	38	100.0

Table 4. System of successful queries in combined System B Sessions (Legend: The table lists the systems that have provided documents with a relevance score of 2 or 3 in combined system B sessions. That explains why N=38.)

Table 5 expands on Table 4 and presents the share of successes at query level. Table 5 presents all queries carried out in system B, both distinct system B queries and combined system B queries. Although the participants in a number of cases found the categorization irrelevant, it was still used in approximately two thirds of the queries (see outer right hand column). In addition, when calculated in terms of the share of successful queries, queries including categories had a better performance (24.2% of queries were successful) than queries omitting categorization (16.7% of queries were successful). To sum up, in combined system B searches, more than half of system B sessions included system A queries to some extent. However, at the query level for all system B queries, queries including a category had a larger chance of succeeding compared to queries that basically corresponded to system A queries.

	Success	Failure	Total
Queries with categories	52 (24.2)	163 (75.8)	215 (100.0)
Queries without categories	20 (16.7)	100 (83.3)	120 (100.0)
Total	72	263	335

Table 5. System B Queries: Frequency of Category Use and Query Success (Percentages) (Legend: The table contains all queries processed in system B, both regular system B queries and combined system B queries (N=335).)

In the post search interviews, participants were asked to assess system B. In the responses, we found answers to when the categorization was useful and when it was not. The answers are analysed in this section in order to elaborate further on the results gained from the search log presented above. There was an overall agreement among the participants that the categorization was useful when they had a large set of results. P21 discussed a query with 14 results:

> It did not help me so much there because the query didn't have that many results. It was possible to cope with the documents there, whether the categorization had been there or not. Only 14 documents were retrieved. You could cope with that. It is [more] helpful when you get large results, a thousand documents or so (P21).

When the categorization was useful in terms of retrieval, set sizes varied. Some mentioned 40 documents, others like P21 mentioned far more. Categorization was also found useful in generating new perspectives on the composition of a query and for understanding the facets of the search task. That supports the decision of coding combined system B queries and sessions as system B queries and sessions in the overall coding of the search log. One example was given by P02, who would have liked to have access to the categorization in a system A session:

> At the end I would have liked to be able to go over there [into the categorization], because no matter what I did, I could not find anything. And then I need somewhere else to search where I have the option of seeing other sub-topics in order to perhaps access it that way (P02).

P09 supports this statement when discussing a system B session:

> It worked well there, because suddenly I found a principal topic that I could click on. And that gave me that … Hey! Yes! That has to do with company taxation. So it also helped me thinking what this is at all (P09).

These findings confirm Käki's (2005b) findings that "the original query was vague, broad, general, or contained words that have multiple meanings" (Käki 2005b 138). Still, the present participants discussed whether categorization was more useful to people with some or no insight into the topic of the tasks. P06 knew what to look for in one of the tasks:

> I knew that if I was to look for something about the taxation then I would also know something about independent businesses. And then I could go in there faster. So I knew that I should choose "personal incomes" over "capital income" [examples of categories]. I know the tax rules. So it is easier to choose between the categories when the answer is known in advance (P06).

P20 on the other hand did not find much help from categorization:

> But I don't know if I would ever start going through all this [the categories]. I think it takes more time because I don't know what is behind. If I was a specialist in SKAT and knew all about company tax settlements or the like, then [the categorization] might be perfect for me because then I would know that I can go in there exactly, click that, and get the documents out. But I don't know if it would [omit] some documents that I need, if it limits the results too much (P20).

P24 sums up the usefulness for users with a lot of knowledge of the task topic and users with less knowledge:

> If I know what I am looking for, or at least think I know where to go [in the categories], then it is really good. But when I don't know, it might also be good because you get to try out different keywords [taxonomy terms]. But if you have the wrong keyword, you will definitely not find it that way (P24).

The reason for the difference of opinion may be due to lack of insight into system functionalities and taxonomy. Thus, a considerable number of the participants mentioned lack of experience with the test system as an important reason for difficulties experienced in locating relevant documents. The difficulties can be seen in Table 2. Here 34% of all system B reformulations consist of changing the category, meaning that participants clicked around between categories without changing the remainder of the search options. In other cases, the trouble experienced by the participants was caused by apparently curious categorizations offered by system B. One example was the presence of the taxonomy term "tonnage taxes" in a query regarding property gain taxes (P13). We have already mentioned the varying sizes of the documents in the collection and the importance of giving employees directions regarding document type. The findings suggest that in collections with large documents, the documents should be indexed in smaller units to obtain more precise search results. On the other hand, when using categorization in search results that are already very limited, as was the case in many system B searches, the results may be skewed. This may be due to lack of experience with the categorization in system B, too narrow queries or odd suggestions for categories. These reasons may explain the increased number of queries in system B sessions. P14 summarises the discussion by saying:

> Once you begin to get an idea [of] what the categories are, what they stand for … then you fumble until you find out what it is. Are there more roads leading to Rome, or which is the fastest, or …? Well, it is an adaptation with some things. What is the wisest thing to do (P14).

5.0 Limitations

We recognize that the search test has limitations. The test was methodologically challenged by the preliminary state of the test database. A running intranet might have generated different performance measures and searching behaviour among the participants. Also, we investigated the information searching of a large institution with highly specialized employees. We may not be able to apply the findings in smaller governments with generalist employees. However, the search test represents a user based and realistic evaluation of automated categorization, which adds to the limited body of knowledge within specialized e-government retrieval and indexing.

6.0 Conclusions

With the present paper we wanted to investigate the comparative performance of free text indexing (system A) and automated categorization (system B). The purpose of the study was to identify and characterize the potential role of categorization in professional e-government information retrieval. We found that free text indexing outperforms categorization when compared in terms of quantitative

measures such as the number of reformulations, session success, and query success. Different causes were found for the increased effort to retrieve relevant documents in system B. Examples are trouble finding suitable categories due to lack of knowledge of the taxonomy. The taxonomy challenges were also identified in the analysis of types of reformulations in system B, where many reformulations consisted of a change of category alone. In relation to retrieval system design the results stress the importance of an appropriate and meaningful level of detail in controlled vocabularies. From the interviews we found qualitative explanations to the potential of categorization despite the differences in performance between the two systems. We found that categorization was useful: 1) if the query retrieved large sets of results; 2) in suggesting new search words for a query; and, 3) for understanding the facets of a search. On the other hand, categorization was not useful if: 1) a highly relevant result came out among the first results; or, 2) if the set of results turned out to be very small. Overall, it is concluded that there is a basis for implementing categorization in information systems supporting professional e-government users. Categorization is a valuable component in successful retrieval in the domain too to support everyday information needs. Therefore we recommend applying categorization in e-government in combination with other search features to back different types of information needs among employees.

References

Apté, Chidanand, Damerau, Fred and Weiss, Sholom M. 1994. Automated learning of decision rules for text categorization. *ACM transactions on information systems* 12 no. 3: 233-51.

Borlund, Pia. 2003. The IIR evaluation model: A framework for evaluation of interactive information retrieval systems. *Information research* 8 no. 3. Available http://www.informationr.net/ir/8-3/paper152.html.

Carpineto, Claudio, Osiński, Stanislaw, Romano, Giovanni and Weiss, Dawid. 2009. A survey of web clustering engines. *ACM computing surveys* 41 no. 3: 17:1-17:38.

Chen, Hsinchun. 1995. Machine learning for information retrieval: Neural networks, symbolic learning, and genetic algorithms. *Journal of the American Society for Information Science* 46: 194-216.

Choo, Chun Wei. 2006. *The knowing organization: How organizations use information to construct meaning, create knowledge, and make decisions,* 2nd ed. New York: Oxford University Press.

Chung, EunKyung, Miksa, Shawne and Hastings, Samantha K.. 2010. A framework of automatic subject term assignment for text categorization: An indexing conception-based approach. *Journal of the American Society for Information Science and Technology* 61: 688-99.

Dumais, Susan, Platt, John, Heckerman, David and Sahami, Mehran. 1998. Inductive learning algorithms and representations for text categorization. In Makki, K. and Bouganim, L., eds., *CIKM '98 Proceedings of the seventh international conference on Information and knowledge management*. New York: ACM, pp. 145-55.

Edmunds, Angela and Morris, Ann. 2000. The problem of information overload in business organisations: A review of the literature. *International journal of information management* 20: 17-28.

Fang, Zhiyuan. 2002. E-government in digital era: Concept, practice, and development. International *Journal of the computer, the Internet and management* 10 no. 2: 1-22.

Goh, Dion Hoe-Lian, Chua, Alton Yeow-Kuan, Luyt, Brendan and Lee, Chei Sian. 2008. Knowledge access, creation and transfer in e-government portals. *Online information review* 32: 348-69.

Grant, Gerald and Chau, Derek. 2005. Developing a generic framework for e-government. *Journal of global information management* 13 no. 1: 1-30.

Jaeger, Paul T. 2003. The endless wire: E-government as global phenomenon. *Government information quarterly* 20: 323-31.

de Jong, Menno and Lentz, Leo. 2006. Municipalities on the web: User-friendliness of government information on the Internet. In Wimmer, Maria A., Andersen, Kim Viborg, Grönlund, Åke and Scholl, Hans J., eds., *Electronic government: 5th International Conference, EGOV 2006, Kraków, Poland, September 4-8, 2006. Proceedings*. Berlin: Springer, pp. 174-85.

Kettunen, Kimmo and Henttonen, Pekka. 2010. Missing in action? Content of records management metadata in real life. *Library & information science research* 32: 43-52.

Klischewski, Ralf. 2006. Ontologies for e-document management in public administration. *Business process management journal* 12: 34-47.

Kopackova, Hana, Michalek, Karel and Cejna, Karel. 2010. Accessibility and findability of local e-government websites in the Czech Republic. *Universal access in the information society* 9: 51-61.

Koshman, Sherry, Spink, Amanda and Jansen, Bernard J. 2006. Web searching on the Vivisimo search engine. *Journal of the American Society for Information Science and Technology* 57: 1875-87.

Kraemer, Kenneth L. and Dedrick, Jason. 1997. Computing and public organizations. *Journal of public administration research and theory* 7: 89-112.

Kules, Bill and Shneiderman, Ben. 2004. Categorized graphical overviews for web search results: An exploratory study using U. S. government agencies as a meaningful and stable structure. In *Proceedings of the Third*

Annual Workshop on HCI Research in MIS. Washington, D.C.: AIS SIGHCI, pp. 20-3.

Käki, Mika. 2005a. *Enhancing web search result access with automatic categorization*. Ph.D. Tampere: Department of Computer Sciences, University of Tampere.

Käki, Mika. 2005b. Findex: Search result categories help users when document ranking fails. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. Portland, Oregon: ACM, 131-40.

Käki, Mika and Aula, Anne. 2005. Findex: Improving search result use through automatic filtering categories. *Interacting with computers* 17: 187-206.

Lykke, Marianne, Price, Susan and Delcambre, Lois. 2012. How doctors search: A study of query behaviour and the impact on search results. *Information processing & management* 48: 1151-70.

Moen, William. E. 2001. The metadata approach to accessing government information. *Government information quarterly* 18: 155-65.

Qi, Xiaoguang and Davison, Brian D. 2009. Web page classification: Features and algorithms. ACM *computing surveys* 41 no. 2: 1-31.

Qu, Bo, Cong, Gao, Li, Cuiping, Sun, Aixin and Chen, Hong. 2012. An evaluation of classification models for question topic categorization. *Journal of the American Society for Information Science and Technology* 63: 889-903.

Sabucedo, Luis Álvarez and Rifón, Luis Anido. 2006. Semantic service oriented architectures for eGovernment platforms. *American Association for Artificial Intelligence*. Available http://www.aaai.org/Papers/Symposia/Spring/2006/SS-06-06/SS06-06-018.pdf.

Schwartz, David G., Divitini, Monica and Brasethvik, Terje. 2000. *Internet-based organizational memory and knowledge management*. Hershey, USA: Idea Group.

Sebastiani, Fabrizio. 2002. Machine learning in automated text categorization. *ACM computing surveys* 34 no. 1: 1-47.

Sormunen, Eero. 2002. Liberal relevance criteria of TREC – counting on negligible documents? In Hancock-Beaulieu, Micheline, ed*., SIGIR 2002 : proceedings of the Twenty-Fifth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 11-15, 2002, Tampere, Finland*. New York: Association for Computing Machinery, pp. 324-30.

Sprehe, J.Timothy, McClure, Charles R. and Zellner, Philip. 2002. The role of situational factors in managing U.S. federal recordkeeping. *Government information quarterly* 19: 289-305.

Tambouris, Efthimios, Manouselis, Nikos and Costopoulou, Constantina. 2007. Metadata for digital collections of e-government resources. *The electronic library* 25: 176-92.

The Danish Government, Local Government Denmark (LGDK) & Danish Regions. 2007. *The Danish e-government strategy 2007-2010: Towards better digital service, increased efficiency and stronger collaboration*. Available http://www.modernisering.dk/fileadmin/user_upload/documents/Projekter/digitaliseringsstrategi/Danish_E-government_strategy_2007-2010.pdf.

Turmo, Jordi, Ageno, Alicia and Català, Neus. 2006. Adaptive information extraction. *ACM computing surveys* 38 no. 2: 4

Zamir, Oren and Etzioni, Oren. 1999. Grouper: A dynamic clustering interface to web search results. *Computer networks: The international journal of computer and telecommunications networking* 31: 1361-74.