The Role of Automated Categorisation in e-Government Information Retrieval

Tanja Svarre & Marianne Lykke

Dept. of Communication & Psychology, Aalborg University, Aalborg, Denmark, tanjasj@hum.aau.dk

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, navigation and metadata. This paper investigates how automated categorisation can enhance information organisation and retrieval and presents the results of a realistic evaluation that compared automated categorisation and free text indexing of the government intranet used by Danish tax authorities. Thirty-two individuals participated in the evaluation, which conducted a combination of simulated searches and genuine search tasks. Searching behaviour and search outcome was documented by search logs, relevance assessments, and post search interviews.

The evaluation demonstrates a high potential for automated categorisation in a government context. Overall, the categorised organisation generated more reformulations at session level and less success at query level. Session success was found to be fairly even between the two systems. Categorised overviews were useful if the participant did not possess much knowledge of the task at hand. When task knowledge was present, categorisation was used to support the assumptions of a correct search. Participants avoided automated categorisation if high-precision documents were among the top results or if few documents were retrieved. The findings emphasise the importance of simultaneously providing different search options for e-government IR systems, and they reveal that automated categorisation is a valuable candidate for improving search facilities within this domain.

1. Introduction

e-Government facilitates governments utilising ICT to communicate with and allow access to information for external parties (e.g. Fang, 2002; Jaeger, 2003; Grant & Chau, 2005). Documental support is a key issue for operations undertaken in public administrations (Kraemer & Dedrick, 1997; Klischewski, 2006; Sabucedo & Rifón, 2006). Therefore, not being able to find needed information can have severe human *and* financial costs (Kraemer & Dedrick, 1997). The combination of the importance of information and information overload has the potential to produce undesirable situations for e-government stakeholders. Furthermore, studies indicate that the facilities of e-government systems still leave room for improvement, for instance, in terms of searching (e.g. Goh et al., 2008), navigation (e.g. de Jong & Lentz, 2006) and the extent of metadata adoption (e.g. Kopackova, Michalek & Cejna, 2010).

Edmunds & Morris (2000) mention different methods to reduce information overload in organisations, among which is value-added information. A concrete way of adding value to information is through the assignment of metadata. The assignment of metadata in the domain serves several purposes, for instance, allowing interoperability between systems and enabling users to retrieve more precise search results (Moen, 2001; Tambouris, Manouselis & Costopoulou, 2007). Metadata ease knowledge sharing between employees in e-government, within organisations and externally (Schwartz, Divitini & Brasethvik, 2000; Choo, 2006). Metadata can be assigned either manually by humans or automatically on the basis of a machine generated analysis of the words constituting the documents. In e-government, the predominant approach is manual assignment.

In the field of US federal records management, Sprehe, McClure and Zellner (2002) found, that different situational factors affected the quality of federal employees' record keeping, causing a divergence in the quality of the 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 record

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 & Henttonen, 2010). The results suggest that e-government indexing might 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 organisations (e.g. Choo, 2006). If automatic indexing can improve subject metadata, then there is reason to assume that the retrieval and sharing of knowledge in the domain is also influenced in a positive sense.

2. Categorisation

Categorisation places documents in categories, usually in a web based environment, with the purpose of supporting searches (Qi & Davison, 2009). Specifically, categorisation enables post-limitation of search results on the basis of document characteristics, e.g. subject, document type, authors, etc. Categorisation may be based on either manually added metadata or automated procedures. Automated procedures count 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 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 categorisation of the full collection of documents (Sebastiani, 2002).

Automated categorisation 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 count Apté, Damerau and Weiss (1994), Chen (1995) and Dumais et al. (1998). Turmo, Ageno and Català (2006), Chung, Miksa and Hastings (2010), and Qu et al. (2012) are more recent examples. User-based evaluations are rarer. Zamir and Etzioni (1999) have evaluated their cluster based interface Grouper by means of search logs. They found that users explored several clusters to locate relevant documents and that the Grouper users found more documents compared to the baseline system (named HuskySearch). However, as the log was not supplemented by qualitative data, the study does not provide explanations for the identified differences between the test system and the baseline system.

In 2004, (Kules & Shneiderman) made a comparative study of ranked and categorised outputs. The background of the participants is not known, but the domain is U.S. government webpages, and the controlled tasks applied were not aimed at people with specialist knowledge of the domain. The tasks were controlled to provide approximately 200 documents for each search. On the basis of the descriptions of the paper, it is not possible to deduce whether the categorisation is based on extracted or assigned indexing. The authors get positive feedback from the participants. They 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 categorisation throughout the test despite the limited time available. Despite the controlled character of the test, the authors conclude that categorisation is highly useful in supporting understanding of large sets of search results.

A later study based on extracted indexing is Käki's (2005a; 2005b; Käki & Aula, 2005) investigation of categorisation of web documents. Two algorithms for extracting category candidates were applied. On the basis of the two extraction algorithms, two interfaces were set up for testing. Different evaluations have been reported from the study. Käki and Aula (2005) made a comparative study of an interface comprising the algorithm and categorised search interface with the World Wide Web as the test base. Twenty test

persons and nine predefined queries in general topic areas formed the basis of the test. The study found that the categorised interface had a better average performance in precision (62% against 49%) *and* recall (33% against 19%). The results of the test persons' attitudes against the two systems demonstrated a more positive attitude towards the test system compared to the baseline system. The test was followed up by a three-month longitudinal study with 16 participants (Käki, 2005b). The participants did not receive any instruction on the use of the test system besides using it any way they would like. The study 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, p. 138). The ability of the categories to help increase the focus of a less precise query was also expressed in the second questionnaire. Furthermore, categories were found useful when result rankings were deficient. The results of the study are interesting because they demonstrate that categorising results is not necessarily useful in all information searching situations. From the analysis, we do get an indication of situations in which categories may be useful. However, a more systematic investigation would be relevant.

Many studies have examined automatic categorisation on the basis of various techniques but few were conducted with the participation of users. In the present paper, we are investigating categorisation on a corporate and e-government intranet by including professional users within that domain. On this basis, the research questions guiding our further work run as follows:

1) How does full text indexing and automatic categorisation perform in relation to:
   a) Number of queries in sessions?
   b) Number of terms in queries?
   c) Number of concepts in queries?
   d) The type of search operator applied?
   e) The use of document type filters?
   f) Number of reformulations?
   g) Types of reformulations?
   h) Degree of search success in queries and sessions?
   i) Overall performance measured by performance measures?
2) What are the implications for the adoption of categorization in e-government retrieval systems?

3. Methodology

For answering the research questions, we carried out a search test in a realistic setting in a real life government intranet at the Danish Tax Corporation, SKAT. The prevailing task of SKAT is the collection of taxes in Denmark. The organisation handles administration related to taxes, duties, customs, debt collection, tax assessment of real estate and cars and gaming activities. The organisation has approximately 7,200 employees located at different office locations across Denmark (SKAT, 2012). The test took place in June 2010 in two office locations of SKAT. The running intranet of the organisation 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 organisation and departments. At the time of the test, the intranet contained 681,640 documents.

The search test compares full text indexing, (extracted indexing, system A) and categorisation (assigned indexing, system B) in an experimental manner. A prototype of the organisation's future intranet functioned as the test system (system B) of the search test. The test system contained a random sample of the running intranet. The sample comprised 188.600 documents corresponding to 28% of the full document collection. The prototype was based on CMS technology. Autonomy's (www.autonomy.com) search software, IDOL, provided the search functionalities of the search interface. In IDOL,

categorisation is based on machine learning. The taxonomy used for the categorisation has 169 terms divided into two levels. The interface of the prototype is depicted in Figure 1. Though more fields were available, the participants only used the search fields query box, search operator and document type during testing. The test system facilitated four types of searches: free text (best match), all the words (Boolean AND), this exact sentence (phrase search) and at least one of the words (Boolean OR). Search results were relevance ranked. For each hit, the document title, a snippet highlighting the search terms and the surrounding terms, the document type and the date of publication appeared. The categorisation (system B) is shown in Figure 2 (the box at the right hand side of the result list). The selection of one or more categories took place after a search had been carried out and a result existed. On the basis of the retrieved documents, the search result was limited to subjects present in the search results. The categorisation 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 right hand side of the screen 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.

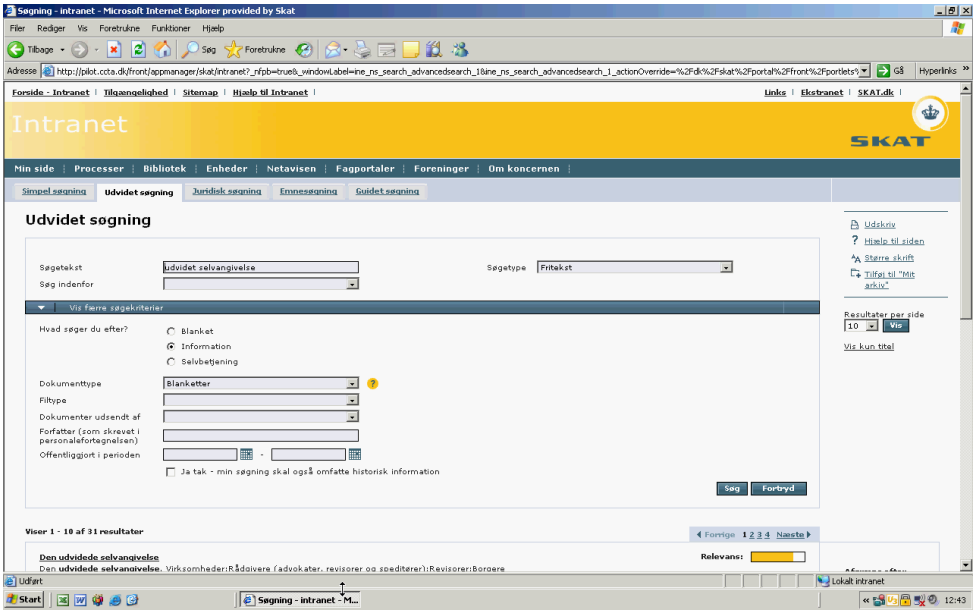

Figure 1 Screen dump of the test system: search fields

Thirty-two employees participated in the test. The participants were recruited using a questionnaire distributed by e-mail. In our selection of participants, we emphasised frequency of intranet use and of information seeking. Forty-two employees met the requirements. Of these, 10 were used as pilot testers, while the remaining 32 participated in the actual test. We employed three simulated and one genuine work task in the test. 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 three parts: (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 could not try out the prototype ahead of the test. In all test sessions, both the succession of tasks and systems were rotated. When searching in system B, the participants were obliged to use categorisation 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 of the test, a short post search interview was conducted. The test sessions ranged between 30 minutes and two hours. The test setting (recruitment questionnaire, search tasks and the general test session) was pilot tested ahead of data collection.

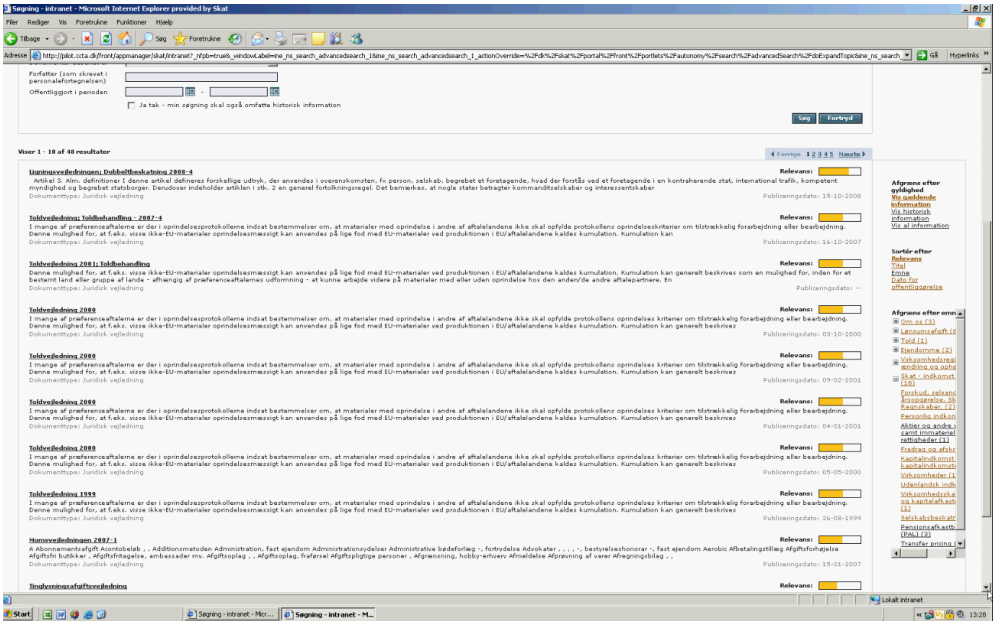

Figure 2 Screen dump of the test system: categorised overview

Different data was collected throughout the search test. The participants' interaction with the test system was logged using the software Morae (see http://www.techsmith.com/morae.asp). Interviews, both oral and in questionnaire form, were carried out along during the course of the search test. The recruitment questionnaire provided background data on the participants' demographic data, seeking behaviour, etc. The relevance of retrieved documents was assessed through the test. Relevance was assessed on a four-point scale reflecting Sormunen's (2002) proposed four-point scale. Additionally, during the search test, the simulated work tasks were assessed on the basis of the participants' knowledge of the subject, their perception of the degree of difficulty and the extent of similarity with their genuine work tasks. Finally, after the search tasks, a post interview was carried out. For documentation purposes, a Dictaphone was set to record the search test and the post interview. The recordings were subsequently transcribed.

The data collected consisted of (1) background data (from the recruitment questionnaire), (2) interview transcriptions (from the search sessions and the post search interview), (3) search logs and (4) relevance assessments. Background data were analysed using descriptive statistics. We used atlas.ti for analysing the interviews after transcription. The search log registered search time and keys 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. Finally, the relevance assessments of documents were typed into SPSS. The sum of this work resulted in the identification of a number of variables listed in Table 1. After the registration of data, statistical analyses

were carried out. The analysis consisted of univariate and bivariate statistics, frequencies, means and correlations.

Table 1. Variables Identified in the Data Set

Variable	Definition	Measurement
Query level		
Terms per query	Number of words separated by a single spacing. Dashes were not counted as single terms. Terms connected with a dash (e.g. 'e-commerce') were counted as one term.	Average number of terms per query
Use of search operators in queries	The search operator chosen for a specific query	Distribution of queries using each of the four search types in percentages
Use of the filter 'Document type' (DT) in queries	The DT filter chosen (if any) for a specific query	Average number of queries using the DT filter in percentages
Number of hits in queries	The number of hits retrieved in queries.	Average number of hits retrieved
Query success	Queries retrieving at least one document with a relevance score of 2 or 3 are considered successful	Percentage of successful queries
Type of reformulations	Reformulations in queries (registered as the change from the past to the present query; types include category, query terms, document type, search operator, and a combination of the above)	Percentage of reformulations in queries
Session level		
Number of sessions with reformulations	Number of sessions containing more than one query; reformulations comprise changes of queries, search type (or categories in system B) or document type	Percentage of sessions with reformulations
Number of reformulations per session	Number of times a query has been reformulated in a session	Average number of reformulations per session
Session success	Sessions containing at least one successful query are considered successful	Average number of sessions solved

4. 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 2 summarises the general findings. The average number of terms used in the queries of the test is 2.25 for system A and 2.43 for system B. This corresponds to the average number of terms found in similar studies. For instance Jansen, Spink and Saracevic (2000, p. 214) measured an average of 2.21 terms in their analysis of search logs in Excite. In a log analysis of a university OPAC, Lau and Goh (2006, p. 1322) found the average query length to be 2.86. In a clustering search engine (vivisimo.com) also based on log analysis, Koshman, Spink and Jansen (2006, p. 1879) found and average of 3.13. Some years later, Hochstotter and Koch (2009, p. 55) identified a slightly lower average (between 1.6 and 1.8) in their study based on live tickers in a number of general and metasearch engines. Lately, Lykke, Price and Delcambre (2012) found averages of 1.5 and 2.0 in their comparative search test of a web-based health portal. Finally, in a study comparing categorised searches with non-categorised searches, Käki (2005b, p. 136) found an average of 2.10 for the former and 2.04 for the latter. Our findings corresponds to the findings of similar studies and supports that on average more search terms are applied in categorised queries than in non-categorised queries. The slightly higher average of terms in system B is reflected in the average number of search keys. Thus, system B queries contain 1.90 search

keys compared to 1.67 in system A. To compare, the differences between average number of terms and search keys in Lykke, Price and Delcambre's (2012) study was slightly lower compared to the present results. Thus, the participants used more terms to represent search keys in the present evaluation. On the basis of the data it cannot be deduced, whether the difference is caused by the participants' insight into the search tasks, the nature of the tasks or other factors.

Table 2. General Findings of Variables in Search Test

Variables	System A Sessions N=64 Queries N=229	System B Sessions N=64 Queries N=335
Number of terms in queries (averages)	2.25	2.43
Search filter 'document type' applied (percentages)	43.2	31.6
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

Both systems offered filtering by document type. The filter was used in 42.3% of queries in system A and in 31.6% of queries in system B. This distribution was expected, as system A has fewer query specification options. 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 obviously 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, Price and Delcambre (2012) found an average of 2.5 and 3.2 queries per session. Koshman, Spink and Jansen's (2006, p. 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.

4.1 Sessions and queries

The success of sessions and queries is summed up in Table 3.. The total success at session level slightly benefits system A, with relevant documents found in 89.1% of all sessions. System B succeeded in 84.4 sessions. A specification as to search tasks reveals a fairly even distribution of successful sessions between the two systems except in sim2. For the remainder of the sessions, the systems performed equally, and even had a minor advantage for system B. In sim2, one session failed in system A, while seven sessions failed in system B. At query level, the total number of successful searches is fairly even between the two systems; however, in system B, the total numbers of failed queries are markedly higher than in system A, particularly those in sim1 and sim2, and when compared at a system level (see the last two columns in Table 3). Thus, the performance at query level increases the differences of performance to system A's advantage compared to the more even performance at the overall session level. In short, the two systems provide approximately the same number of successful queries, but there are more failed queries in system B.

The overall comparison of the two test systems shows a slight advantage of system A at session level in terms of ability to retrieve relevant documents. The advantage of system A increases when measured at query level. In addition, system A differs from system B, as fewer terms are needed in queries and the share and number of reformulations is lower. In the next sections, we will explore the nature and causes

of the difference of performance of the two systems. We will explore what characterises the search situation, the number and types of reformulations carried out and the unintended use of system A in system B searches.

Table 3, Session and Query Success (Percentages)

	Sim1		Sim2		Sim3		NWT		Total	
	SysA	SysB	SysA	SysB	SysA	SysB	SysA	SysB	SysA	SysB
Session succeeded	15 (93.8)	16 (100.0)	15 (93.8)	9 (56.3)	16 (100.0	16 (100.0	11 (68.8)	13 (81.3)	57 (89.1)	54 (84.4)
Query succeeded	18 (58.1)	23 (33.3)	17 (30.4)	11 (9.7)	20 (27.8)	22 (25.6)	15 (21.4)	16 (23.9)	70 (30.6)	72 (21.5)

As appears from the total numbers, more queries are executed in system B than in system A. This is also the case at task level (see Table 4). Here, the average number of queries needed in order to solve a task in system differs with almost two queries (the last column). With regards to the individual search tasks, the genuine information need has a slightly lower average in system B compared to system A, indicating that genuine information actually benefitted from the categories. In the remainder of the search tasks, system B is higher than system A in terms of averages. It has already been shown that sim1 and sim2 contained a significantly higher share of failed queries in system B compared to system A. Table 4 shows where sim1 and sim2 executed in system B have an average of queries twice as large as system A.

Table 4. Number of Queries in Sessions at Task Level (Averages)

	Sim1	Sim2	Sim3	NWT	Total
System A	1.94 (n=16)	3.50 (n=16)	4.50 (n=16)	4.38 (n=16)	3.58 (n=64)
System B	4.31 (n=16)	7.06 (n=16)	5.38 (n=16)	4.19 (n=16)	5.23 (n=64)
Total	3.13 (n=32)	5.28 (n=32)	4.94 (n=32)	4.28 (n=32)	4.41 (n=128)

The equivalent table at query level reveals that the number of search terms applied at task level varies between the tasks (see Table 5). Table 5 shows that, except for sim3, more terms have been entered in all system B queries compared to system A. Here, the average number of terms is notably lower in system B than in system A. Due to the scores of sim3, the connection between the average numbers of search terms in queries is not consistently higher in system B than in system A.

Table 5. Number of Search Terms in Queries (Averages)

	Sim1	Sim2	Sim3	NWT	Total
System A	2.32 (n=31)	2.39 (n=56)	2.42 (n=72)	1.94 (n=70)	2.25(N=229)
System B	2.54 (n=69)	2.88 (n=113)	1.79 (n=86)	2.39 (n=67)	2.43 (N=335)
Total	2.47 (n=100)	2.72 (n=169)	2.08 (n=158)	2.16 (n=137)	2.36 (N=564)

4.2 Search operators

In the search interface, the default setting of the search operator field is the best match operator (BM). Therefore, as the participants did not have any prior experience with the test systems, it was plausible that the default BM operator would be most frequently used in the test. Thus, users tended to use the default settings put forward by the system (Markey, 2007a, p. 1077). As expected, the BM operator had a high frequency across the queries, along with the Boolean AND operator. System A has a slightly higher frequency of BM searches, while the opposite is the case for system B, which was unexpected. Thus, in

system B, the AND is more frequently used than BM (see Table 6). The combination of the more restrictive AND and the mandatory categorisation in system B it is likely to result in large differences between the sizes of search results in the two systems. One explanation for the unexpected distribution of BM and AND operators between system A and system B is that some participants had trouble incorporating the two operators and separating them from each other. Thus, participants intermittently wondered why search terms did not occur in their result list when using the BM operator, for example:

'Yes, but on the other hand it could also give… [best match]… then they all ought to come…' (P15)

In addition, the participants consistently used more search terms when applying the AND operator rather than the BM operator, which resulted in a gap in the number of documents retrieved. To illustrate, an average search in system A using BM retrieved 548 documents, while AND in the same system on average retrieved 121 documents. In system B, average BM searches retrieved 25 documents, while average AND searches retrieved 10 documents (see Svarre, 2012). Thus, the searches carried out in system B were significantly narrower than the broader system A searches, as the search results were also filtered according to subject. In terms of Boolean logic, the addition of a category corresponds to combining a query with an additional term, and, in some cases, an additional concept.

Table 6. Distribution of Search Operators in Queries and Search Terms Used with Operators

	Total distribution of search operators (percentages)		Total number of search terms applied with search operators (averages)	
	SysA	SysB	SysA	SysB
Best match	102 (44.5)	177 (52.8)	1.94 (n=102)	2.13 (n=177)
Boolean AND	110 (48.0)	145 (43.3)	2.51 (n=110)	2.74 (n=145)
Phrase search	13 (5.7)	11 (3.3)	2.08 (n=13)	3.27 (n=11)
Boolean OR	4 (1.7)	2 (0.6)	3.75 (n=4)	2.0 (n=2)
Total	229	335	229	335

Again, it appears that some participants had trouble fully grasping the comparative implications of the two search operators. It has not been possible to deduce causes for the difference in search operators between the two systems in the search interviews, as the participants did not address it during their conversations. That the understanding of Boolean operators challenges end users corresponds to the findings of similar studies (eg. Markey, 2007a).

Table 7. Success of Search Operators (Percentages)

	System A				System B			
	BM	AND	Phrase	OR	BM	AND	Phrase	OR
Success	33 (32.4)	30 (27.3)	7 (53.8)	0 (0.0)	38 (21.5)	32 (22.1)	1 (9.1)	1 (50.0)
Total	102 (100.0)	110 (100.0)	13 (100.0)	4 (100.0)	177 (100.0)	145 (100.0)	11 (100.0)	2 (100.0)

Legend: BM=Best match, AND=Boolean AND, Phrase=Phrase search, and OR=Boolean OR.

The use of search operators has resulted in significant differences in the number of hits retrieved in the two test systems due to the use of search operators. However, the success rate of the queries in terms of search operators is needed to identify which performed better (see Table 7). The success rate of system A is higher on a general level compared to system B. System A queries have a slightly higher success rate in BM searches compared to AND searches.

To conclude, the best performance was found in system A using the FT operator. This also indicates well-functioning relevance ranking within the system. Furthermore, system A managed to perform better on the basis of the broader queries applied. The participants had difficulty applying and understanding the search operators correctly, which led to a weaker performance of system B due to a majority of small result sets.

4.3 Reformulations

Table 8. Number of Sessions with Query Reformulations (Percentages)

	Sim1		Sim2		Sim3		NWT		Total	
	SysA	SysB	SysA	SysB	SysA	SysB	SysA	SysB	SysA	SysB
Reformu-lations	6 (37.5)	11 (68.8)	12 (75.0)	16 (100.0)	10 (62.5)	15 (93.8)	14 (87.5)	11 (68.8)	42 (65.6)	53 (82.8)
Total	16	16	16	16	16	16	16	16	64	64

Reformulations provide information about whether and how searchers try to correct a query on the basis of an unsatisfying search result. Table 8 and Table 9 specify the reformulation figures. From the Table 8, it appears that the general figures are mirrored at session level with one exception. The genuine search task is the only task with fewer reformulations in system B than in system A. However, the number of reformulations is still high. In sessions with reformulations, the average number of reformulations was 4.59, a little less for system A searches and a little more for system B searches (see Table 9, bottom-right cell). From Table 9, it is apparent that sim3 had a higher average number of reformulations in system A. For the remainder of the tasks, system B had the highest average number of reformulations.

Table 9. Number of Reformulations in Sessions (Averages)

	Sim1	Sim2	Sim3	NWT	Total
SysA	2.50 (n=6)	3.33 (n=12)	5.60 (n=10)	3.86 (n=14)	3.93 (n=42)
SysB	4.82 (n=11)	6.06 (n=16)	4.67 (n=15)	4.64 (n=11)	5.11 (n=53)
Total	4.00 (n=17)	4.89 (n=28)	5.04 (n=25)	4.20 (n=25)	4.59 (N=95)

Legend: Sessions without reformulations have been excluded, which makes N=95.

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 terms, the document type or the search operator were changed, if several parameters were changed or if no reformulation occurred (see Table 10). In system A, the overall preferred reformulation is a change of search terms. 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 terms 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 terms 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 terms. This reflects the design of the system, where only categories with content were shown to the searchers. Thus, when search terms 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 terms as a reformulation.

Table 10. Types of Reformulations for All Queries (Percentages)

	Total SysA	SysB
No reformulations	69 (30.1)	62 (18.5)
Category	-	114 (34.0)
Query terms	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)

The success of the respective types of reformulations is summed up in Table 11. Overall, system A has a higher share of successful reformulations than system B. At the level of types of reformulations, the best performance is achieved in system A by using a combination of terms. Here, about 40% of queries manage to retrieve relevant documents. This is followed by a change of settings of the document type filter. In system B, the variance of performance was smaller than in system A, and the participants had less success in improving their outputs by changing query terms and search operators, meaning that the two most frequent reformulation types accounted for roughly the same share of successful queries. Categories, search operators, and a combination of query modifiers had the best performance within the system, but the performance was below the percentages gained in system A. Thus, within system B, we may conclude that reformulations based on a change of categories perform better, and the remainder modification tools System A reformulations were more successful when they consisted of a combination of more parameters. However, the share of successful reformulations leaves room for improvement in both systems.

Table 11. Query Success on the Basis of Types of Reformulations (Percentages)

	System A	Total system A	System B	Total system B
Category	-	-	24 (21.1)	114(100.0)
Query terms	22 (22.7)	97 (100.0)	5 (10.6)	47 (100.0)
Document type	9 (32.1)	28 (100.0)	1 (12.5)	8 (100.0)
Search operators	1 (12.5)	8 (100.0)	1 (20.0)	5 (100.0)
>1 types simultaneously	11 (40.7)	27 (100.0)	19 (19.2)	99 (100.0)

4.4 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 cause was the speed of the system. Thus, in the time waiting for the system to categorise search results, some participants began to review the documents found on the basis of the initial query. On other occasions, the participants actually 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 searches. The following quote 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 categorisation 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. It has previously been observed that system B searches tended to be narrow. 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 categorisation despite the few results, while others omitted the categorisation 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).

Table 12. Sessions Carried Out in System B or in a Combination of System B and System A: Frequency and Success (Percentages)

	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)

Legend: 'System B' denotes sessions that have been carried out in system B exclusively. '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 quote illustrates, in a combined system B search with just two retrieval results, how the participant ends up assessing the documents retrieved without categorisation. This supports the assumption put forward by Kules and Schneiderman (2004, p. 2) that search results must have a certain size to make categorisation useful.

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. Reporting on the extent of combined system B sessions is the purpose of the present section. Table 12 lists the share of combined system B sessions. The table shows that about 60% of the system B sessions contained one or more queries that omitted categories. It is also 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 categorisation are therefore substantial.

Table 13 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 categorisation in some queries of a session, but it may still be that relevant documents are found by means of categorisation.

Table 13. System of Successful Queries in Combined System B Sessions

	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

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 14 expands on Table 13 and presents the share of successes at query level. Table 14 present 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 categorisation 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 categorisation (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.

Table 14. System B Queries: Frequency of Category Use and Query Success (Percentages)

	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

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 categorisation 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 categorisation 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 categorisation 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 categorisation was useful in terms of retrieval, set sizes varied. Some mentioned 40 documents, others like P21 mentioned far more. Categorisation 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 categorisation in a system A session:

'At the end I would have liked to be able to go over there [into the categorisation], *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 (based on extracted categorisation, see above) when 'the original query was vague, broad, general, or contained words that have multiple meanings...' (Käki, 2005b, p. 138). Still, the participants of the present search test discussed whether categorisation 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 categorisation:

'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 categorisation] *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 10. 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 categorisations 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 categorisation 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 categorisation 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. Limitations

We recognize that the search test has limitations. The test was methodically 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. Conclusions

With the present paper we wanted to investigate the comparative performance of full 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. The participants of the test were specialized employees of the tax authorities of Denmark.

We found that system A outperforms system B at a **general level** in terms of the number of terms used, number of reformulations, session success, and query success. However, at **task level** the performance is more even between the two systems, specifically in terms of session success and query success level. Different causes were found for the increased effort to retrieve relevant documents in system B. The more restrictive AND operator was used with the same frequency in both systems, resulting in at times very small result sets in system B. This shows that the participants had difficulties understanding the meaning of the two predominant operators of the system. Further, in system B some participants expressed trouble finding suitable categories in the categorization to match their queries due to lack of knowledge of the taxonomy. The taxonomy challenges were identified in the analysis of types of reformulations in system B too. Here a change of mere categorization accounted for 34% of the reformulations. In relation to design of retrieval systems the results stress the importance of an appropriate and meaningful level of detail in controlled vocabularies.

The participants **avoided using categorization** in one third of the system B queries. Analyses of the queries carried out in system B showed a fairly even distribution of successful sessions as to whether the session had been solved by means of categorization or not. At query level the inclusion of a category was successful in 24.2 % of queries, while of the queries that omitted categories had a success rate of 16.7 %.

Though having a higher success rate, we still needed to understand the nature of system B omissions. Here the post search interviews provided insight. We found that categorization was not supportive in queries, where a highly relevant result came out among the first results. Neither was it relevant, if a very small set of results were retrieved. In those cases the categorization were considered as inconvenient to the retrieval process, as it was easier to manually look through the results instead of deciding on the correct category. On the other hand categorization was useful when a large set of results were retrieved, in suggesting new search terms for a query, and for understanding the facets of a search.

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 in the domain. Therefore we recommend applying categorization in e-government in combination with other search features to maintain different types of information needs among employees.

7. References

Apté, C., Damerau, F. & Weiss, S.M. (1994). Automated learning of decision rules for text categorization. *ACM Transactions on Information Systems, 12*(3), 233-251.

Carpineto, C., Osiński, S., Romano, G. & Weiss, D. (2009). A survey of Web clustering engines. *ACM Computing Surveys, 41*(3).

Chen, H. (1995). Machine learning for information retrieval: Neural networks, symbolic learning, and genetic algorithms. *Journal of the American Society for Information Science, 46*(3), 194-216.

Choo, C.W. (2006). *The Knowing Organization: How Organizations Use Information to Construct Meaning, Create Knowledge, and Make Decisions* (2. ed.). New York: Oxford University Press.

Chung, E., Miksa, S. & Hastings, S.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*(4), 688-699.

de Jong, M. & Lentz, L. (2006). Municipalities on the Web: User-Friendliness of Government Information on the Internet. In: Wimmer, M., Scholl, H., Grönlund, Å. & Andersen, K. (Eds.), *Electronic Government, 5th International Conference, EGOV 2006* (pp. 174-185). Berlin: Springer.

Dumais, S., Platt, J., Heckerman, D. & Sahami, M. (1998). Inductive learning algorithms and representations for text categorization. In: Makki, K. & Bouganim, L. (Eds.), *CIKM '98 Proceedings of the seventh international conference on Information and knowledge management* (pp. 148-155). New York: ACM.

Edmunds, A. & Morris, A. (2000). The problem of information overload in business organisations: a review of the literature. *International Journal of Information Management, 20*(1), 17-28.

Fang, Z. (2002). E-government in digital era: Concept, practice, and development. *International Journal of The Computer, The Internet and Management, 10*(2), 1-22.

Goh, D.H.-L., Chua, A.Y.-K., Luyt, B. & Lee, C.S. (2008). Knowledge access, creation and transfer in e-government portals. *Online information review, 32*(3), 348-369.

Grant, G. & Chau, D. (2005). Developing a generic framework for e-government. *Journal of Global Information Management, 13*(1), 1-30.

Hochstotter, N. & Koch, M. (2009). Standard parameters for searching behaviour in search engines and their empirical evaluation. *Journal of Information Science, 35*(1), 45-65.

Jaeger, P.T. (2003). The endless wire: E-government as global phenomenon. *Government Information Quarterly, 20*, 323-331.

Jansen, B.J., Spink, A. & Saracevic, T. (2000). Real life, real users, and real needs: A study and analysis of user queries on the web. *Information Processing & Management, 36*(2), 207-227.

Kettunen, K. & Henttonen, P. (2010). Missing in action? Content of records management metadata in real life. *Library & Information Science Research, 32*(1), 43-52.

Klischewski, R. (2006). Ontologies for e-document management in public administration. *Business Process Management Journal, 12*(1), 34-47.

Kopackova, H., Michalek, K. & Cejna, K. (2010). Accessibility and findability of local e-government websites in the Czech Republic. *Universal Access In The Information Society, 9*(1), 51-61.

Koshman, S., Spink, A. & Jansen, B.J. (2006). Web searching on the Vivisimo search engine. *Journal of the American Society for Information Science and Technology, 57*(14), 1875-1887.

Kraemer, K.L. & Dedrick, J. (1997). Computing and Public Organizations. *Journal of Public Administration Research and Theory, 7*(1), 89-112.

Kules, B. & Shneiderman, B. (2004). Categorized graphical overviews for web search results: An exploratory study using U. S. government agencies as a meaningful and stable structure, *Proceedings of the Third Annual Workshop on HCI Research in MIS*. Washington, D.C.

Käki, M. (2005a). *Enhancing Web Search Result Access with Automatic Categorization.* Unpublished Doctoral Dissertation, Department of Computer Sciences, University of Tampere, Tampere, Finland, from http://acta.uta.fi/pdf/951-44-6490-7.pdf.

Käki, M. (2005b). Findex: Search result categories help users when document ranking fails. In: *Proceedings of the SIGCHI conference on Human factors in computing systems*, (pp. 131-140). Portland, Oregon: ACM.

Käki, M. & Aula, A. (2005). Findex: Improving search result use through automatic filtering categories. *Interacting with Computers, 17*(2), 187-206.

Lau, E.P. & Goh, D.H.-L. (2006). In search of query patterns: A case study of a university OPAC. *Information Processing & Managament, 42*, 1316-1329.

Lykke, M., Price, S. & Delcambre, L. (2012). How doctors search: A study of query behaviour and the impact on search results. *Information Processing & Managament*(0).

Markey, K. (2007a). Twenty-five years of end-user searching, part 1: Research findings. *Journal of the American Society for Information Science and Technology, 58*(8), 1071-1081.

Moen, W.E. (2001). The metadata approach to accessing government information. *Government Information Quarterly, 18*(3), 155-165.

Qi, X. & Davison, B.D. (2009). Web page classification: Features and algorithms. *ACM Computing Surveys, 41*(2).

Qu, B., Cong, G., Li, C., Sun, A. & Chen, H. (2012). An evaluation of classification models for question topic categorization. *Journal of the American Society for Information Science and Technology, 63*(5), 889-903.

Sabucedo, L.Á. & Rifón, L.A. (2006). *Semantic Service Oriented Architectures for eGovernment Platforms*. Retrieved 08-01-2010.

Schwartz, D.G., Divitini, M. & Brasethvik, T. (2000). *Internet-Based Organizational Memory and Knowledge Management*. Hershey, USA: Idea Group.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys, 34*(1), 1-47.

SKAT (2012). *Ny struktur i Skatteministeriet*. Retrieved 02-07, 2013, from http://skat.dk/getFile.aspx?Id=99397.

Sormunen, E. (2002). Liberal relevance criteria of TREC - counting on negligible documents? In: *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, (pp. 324-330). August 11-15, 2002, Tampere, Finland: ACM.

Sprehe, J.T., McClure, C.R. & Zellner, P. (2002). The role of situational factors in managing U.S. federal recordkeeping. *Government Information Quarterly, 19*(3), 289-305.

Tambouris, E., Manouselis, N. & Costopoulou, C. (2007). Metadata for digital collections of e-government resources. *The Electronic Library, 25*(2), 176-192.

Turmo, J., Ageno, A. & Català, N. (2006). Adaptive information extraction. *ACM Computing Surveys, 38*(2).

Zamir, O. & Etzioni, O. (1999). Grouper: A dynamic clustering interface to Web search results. *Computer Networks, 31*(11-16, 17 May 1999), 1361-1374.