Targeted Query Expansions as a Method for Searching: Mixed Quality Digitized Cultural Heritage Documents

Heikki Keskustalo, University of Tampere, Finland
Kimmo Kettunen, National Library of Finland
Sanna Kumpulainen, University of Tampere, Finland
Nicola Ferro, University of Padova, Italy
Gianmaria Silvello, University of Padova, Italy
Anni Järvelin, University of Tampere, Finland
Jaana Kekäläinen, University of Tampere, Finland
Paavo Arvola, University of Tampere, Finland
Miamaria Saastamoinen, University of Tampere, Finland
Eero Sormunen, University of Tampere, Finland
Kalervo Järvelin, University of Tampere, Finland

Abstract
Digitization of cultural heritage is a huge ongoing effort in many countries. In digitized historical documents, words may occur in different surface forms due to three types of variation - morphological variation, historical variation, and errors in optical character recognition (OCR). Because individual documents may differ significantly from each other regarding the level of such variations, digitized collections may contain documents of mixed quality. Such different types of documents may require different types of retrieval methods. We suggest using targeted query expansions (QE) to access documents in mixed-quality text collections. In QE the user-given search term is replaced by a set of expansion keys (search words); in *targeted* QE the selection of expansion terms is based on the *type of* surface level variation occurring in the particular text searched. We illustrate our approach in a highly inflectional compounding language, Finnish while the variation occur across all natural languages. We report a minimal-scale experiment based on the QE method and discuss the need to support targeted QEs in the search interface.
Keywords: cultural heritage information retrieval, targeted query expansion
Citation: Keskustalo, H., Kettunen, K., Kumpulainen, S., Ferro, N., Silvello, G.,; Järvelin, A., Kekäläinen, J., Arvola, P., Sormunen, E., Järvelin, K., Saastamoinen, M. (2015). Targeted Query Expansions as a Method for Searching Mixed Quality Digitized Cultural Heritage Documents. In *iConference 2015 Proceedings*.

Research Data: In case you want to publish research data please contact the editor.
Contact: heikki.keskustalo@uta.fi, kimmo.kettunen@uta.fi, sanna.kumpulainen@uta.fi, ferro@dei.unipd.it, silvello@dei.unipd.it, anni.jarvelin@uta.fi, jaana.kekalainen@uta.fi, paavo.arvola@uta.fi, eero.sormunen@uta.fi, kalervo.jarvelin@uta.fi, miamaria.saastamoinen@staff.uta.fi

1 Introduction

Digitization of cultural heritage is a huge ongoing effort in many countries. In order to fully disclose the value of Digitized Cultural Heritage (DCH) objects, they should be made smoothly accessible both to general public – students, tourists, amateurs – and to professional users – e.g., historians, archivists and librarians – who may have very different needs and experiences. In order to address different user categories and enhance their engagement with cultural heritage assets, specific and diversified systems and human-computer interaction methodologies should be designed and developed to allow for an intuitive access and exploitation of DCH material (Agosti et al., 2013a; Agosti et al., 2013b).

Moreover, the availability of large DCH material opens up new research opportunities concerning data enrichment, open data access, and close integration with the huge quantity of cultural heritage resources already available in large European digital libraries, such as Europeana. Indeed, DHC documents may contain references to relevant historical events and people which can be exploited for establishing new connections with existing openly available documents (e.g., Europeana collections) which may open up new access paths to the documents and increase discoverability and understandability of the data (Ferro and Silvello, 2014); this aspect is particularly relevant in the context of the Linked Open Data paradigm, which is becoming the new standard for data publishing, discovering and accessing cultural heritage resources (Hyvönen, 2012). DCH collections are composed of historical documents containing compound and inflectional language and variations of words through the years, which can be exploited in diachronic linguistic analyses for studying the evolution of language over time; furthermore, historical

variations of words/sentences extracted from DCH documents could be connected to geographical locations in order to contribute to the development of geo-linguistic atlases used for study syntactic phenomena in languages (Di Buccio et al., 2014).

A fundamental condition to uphold such research opportunities, to develop advanced methodologies and systems for enhancing user engagement and to enrich and analyze DCH resources is being able to automatically process, index, query and retrieve them. DCH may be composed of mixed-quality documents, given that optical character recognition (OCR) techniques may produce recognition errors that affect the index and retrieval techniques and thus also the user experience in accessing and exploiting the data. Mixed-quality documents are due to variability in the quality of historically used fonts (handwriting) and printing technology. In this paper we will address this issue from the query expansion (QE) point of view, which is a relevant building block and powerful component to support the presented vision.

2 Motivation and earlier research

Information retrieval from historical collections is challenging due to string-level variation present in the documents. The variation is due to (i) word inflection; (ii) historical vocabularies; and (iii) noise created by OCR. (Järvelin et al., 2014) Searchers typically prefer using short queries. They consist of only a few search terms (Jansen et al., 2000). In order to cover the string-level variation present in the documents, QE can be performed, in which the user-given search terms are replaced by term variant *sets*. For example, character-level fuzzy matching methods, such as digrams, may be utilized to find fuzzy variants of the search term from the database index (Pirkola et al., 2002; Keskustalo et al., 2003; Järvelin et al., 2014); other methods include use of transformation rules (Pirkola et al., 2003; Loponen et al., 2008); and confusion matrices (Ohta et al, 1997). Alkula (2000) compared performance for Finnish IR with and without lemmatizing and compound splitting; Savoy & Naji (2011) showed how retrieval performance decreases in OCR error corrupted documents. OCR issues are also discussed by Callan et al. (2002), Tong et al. (1996) and Kantor & Voorhees (2000). We will next explain our idea of focused QE based on three dimensions of variation, in order to serve IR in mixed-quality DCH collections.

3 Dimensions of DCH Documents

Next we will briefly describe the types of string-level variation in mixed DCH collections; illustrate the problem of string-level mismatches in such collections; and suggest a solution to this problem. Based on the typology of string-level variation (Järvelin et al., 2014) discussed above, we decided to abstract individual DCH documents as points in a three-dimensional space. The *inflectional* dimension expresses whether the document is made accessible via normalized or non-normalized index words. Normalization may be challenging in digitized DCH collections due to historical vocabularies and high level of noise. The *historical* dimension expresses the extent of archaic word forms occurring in the documents, related to the age of the original documents. The *noise* dimension characterizes the share of character-level errors present in the digitized document. Character errors are produced during the OCR process. We call this space *the DCH Cube*.

From the IR point of view the location of an *individual document* in the DCH Cube is determined by the types of string-level variations in it. We can map any target document into a particular region of space within the Cube by considering the following issues.

1. Inflectional dimension: is the document accessible by normalized words, or not (that is, the index contains inflectional word forms)?
2. Historical dimension: what is the degree of archaic words in the document (how old is the document)?
3. Noise dimension: what is the degree of noise in the digitized document (was OCR used? is the text corrected?)?

Different types of documents are a source of different types of vocabulary mismatches (between the query and the words in the documents) because the extent of inflectional, historical and noisy words may vary considerably among documents. Table 1 illustrates this phenomenon.

Table 1. Examples of string-level mismatches: the user-given query term (in boldface) does not match the various inflectional/archaic/noisy variants which may be present in different types of DCH sub-collections.

Type of variation	Sample term
normalized, modern, error-free	***vuosi*** (“*year*”)
inflectional, modern, error-free	vuoden
normalized, archaic, error-free	wuosi
inflectional, archaic, error-free	wuoden
normalized, modern, noisy	vucsi
inflectional, modern, noisy	vucden
normalized, archaic, noisy	wucsi
inflectional, archaic, noisy	wucden

In Table 1 the search concept “*year*” is expressed by using a normalized modern query term *vuosi*. However, the intended concept may be referred to by a huge number of variants within the DCH Cube. Examples of variants produced by inflection, archaic word forms, noise, and combinations of them are presented. We call these variants *IAN variants*. For example, the index word form *“wucden”* is produced by erroneously scanning the inflectional variant of the archaic word form *“wuoden”* occurring in the original (paper) document.

Focused Query Expansion

We suggest QE based on the region of the Cube in which the intended target document is located. While an individual document could locate in any region, any document should be retrievable. In QE, the original search terms are replaced by a set of its variants. Because DCH collection contains documents of mixed types, we propose constructing *targeted* QEs. In this method different types of QEs are produced to retrieve different types of documents. In more detail, we suggest producing QEs by simulating the “production” of vocabulary mismatches during indexing at any particular region of the Cube. Selected regions of the Cube are described in Table 2.

Table 2. Different document types set demands for different QEs. A sample term *vuosi (“year”)* and some IAN variants for different types of target documents are shown. Noisy words are marked with asterisk (*).

Document type	**Description**	**Examples of QE term variants**
modern, normalized, noise-free	Modern text (proof-read)	*vuosi*
modern, inflectional, noisy	Scanned modern text (OCR)	*vuosi, vuoden, .., vuosl**
historical, inflectional, noise-free	Historical (proof-read)	*wuosi, wuoden, …*
modern+historical, inflectional, noisy	Scanned historical/modern (OCR)	*vuosi, wuosi, wuodcn*, …*

Targeted QE consists of the following steps: (1) the user query terms are used as input for QE; (2) depending on the types of documents (sub-collections) searched, corresponding types of query expansions are produced; (3) in the retrieval phase the properties of the sub-document collections determine which query expansion is used to retrieve documents. In practice, retrieval can be based on sub-collections, which share the same main properties, e.g., whether normalization is used in indexing and whether the documents are noisy or not. Then, the corresponding type of QE will be produced. We will next discuss these ideas in more detail.

4 Methods

4.1 Test collection

Our test collection contains 180,468 digitized historical documents (Finnish newspaper articles from the 1800s) containing archaic vocabulary; 56 topics; and graded relevance assessments. Due to the relevance assessments, relevant documents (the correct answers) for each topic are known in advance. The relevance assessments were made on a four point scale: non-relevant (R0), marginally relevant (R1), relevant (R2), and highly relevant (R3). However, in order to develop the targeted QE principle and start working with the QE software we experimented with a subset of 5 search topics. We made the documents

searchable by via Indri 5.0 search engine[1]. The scanned documents are noisy (see Raitanen, 2012; Järvelin et al., 2014). Kettunen et al. (2014) show that only 13 % of the words of the collection are recognized by a modern Finnish lemmatizer, FINTWOL[2]. A clean, hand-edited word list of the same period gets 58 % of its words recognized by the same lemmatizer. This difference indicates that OCR errors are the main cause of unrecognized words in the collection. Due to noisy word endings we did not apply conflation methods (lemmatization or stemming) but indexed all document words "as-is" (in inflectional and possibly noisy forms).

4.2 Focused QE in Test Collection

Due to test collection properties explained above there is a need to match query term with respect to both modern and archaic word forms; inflectional word forms; and error-free and noisy words. We therefore propose expanding the user query terms along these three dimensions. Finally, after the expansion (producing a set of QE candidates) the term candidates produced are searched from the index of the database. Some of these words actually exist; others do not. The existing terms are ordered by their frequency in the collection - and the top-N words having the highest collection frequency are used as an *expansion set*. For each term its expansion set is structured by using the #synonym operator of *Indri* because the words should be treated as synonyms during the query process.

4.3 Expansion in Three Dimensions

In order to simplify the experiment, we used a minimal subset of rules to modify the user-given terms during query expansion. The query s were given in modern, error-free IAN variants. The rules expressed in Tables 3, 4 and 5 (for inflectional, historical and OCR expansion, correspondingly) were applied.

4.3.1 Inflectional expansion

We used the Frequent Case Generation (FCG) method by Kettunen (2006) to expand the modern word (Table 3). This method produced, for each word, its 22 most common inflectional forms.

Table 3. Examples of inflections created by FCG for a sample term *opetus* ("*teaching"*).

word	inflectional forms (FCG 22)
opetus ("teaching")	*opetuksen, opetusta, opetuksessa, opetuksesta, opetukseen, opetukselle, …*

In Finnish several inflected forms are typically created by adding suffixes (e.g., *opetus -> opetukselle*). OCR errors may affect the suffixes (e.g., *opetukselle -> opetuksellc)* thereby creating invalid word forms. Some searching methods such as fuzzy string matching might miss good variants if they are distant from the query term. Such variants, however, could be produced by combining inflection with regular rules for historical ortography and OCR transformations. Therefore, we started by generating frequent word cases and continue by further modifying these variants by simulating historical changes and OCR noise.

4.3.2 Historical expansion

We next applied regular historical spelling changes. Table 4 presents the subset (one rule) of regular archaic changes used in the experiment.

Table 4. Example of character level variation (v -> w) in old Finnish for sample term *vuosi* ("*year"*)..

modern spelling	historical spelling	modern spelling	historical spelling
v	*w*	*vuosi* ("year")	*wuosi ("year")*

In archaic Finnish "w" was a common letter until the end of the 19th century instead of its modern variant "v". Therefore, we include this particular letter into our automatic letter-transformation rules.

4.3.3 Noise expansion

Table 5 presents a small subset of regular OCR error rules, based on an analyzing the test collection (Raitanen, 2012). In the future, we plan to derive the rules from an independent document sample.

[1] www.lemurproject.org
[2] www2.lingsoft.fi

Table 5. Examples of recognition errors (i -> l) and (t -> l) producing noisy IAN variants.

target letter	error	word form (example)	noisy IAN variants
i	l	kissoille	*klssoille, kissollle*
t	l	diakonissalaitos	*diakonissalailos*

The errors may be created as single character substitutions, deletions and insertions; or fusions or splits between varying numbers of characters (see Ohta et al., 2007). In this study, we replaced (one by one) each of the target letters by an error letter. For example, expanding *"kissoille" ("to cats*") by rule (i -> l) both two "l":s in the word were replaced, thereby producing two noisy QE terms: *"klssoille"* and *"kissollle".*

4.3.4 Test queries

Table 6 explicates the baseline queries and the recall base properties for five sample topics used in the experiment. We used the first four topics of the test collection (#1, #2, #3, #4) plus a fifth topic (#16) entailing letter "v" in order to explore the effects of the historial spelling rule (*v* -> *w*).

Table 6. Finnish baseline queries, their English translations, and the number of documents for each topic, which are judged to be highly relevant (R3), relevant (R2) or marginally relevant (R1).

#	Baseline query	Translation	R3	R2	R1
1	`australia alkuperäiskansa`	Australian aboriginals	27	15	27
2	`diakonissalaitos`	Diacony institution	11	28	90
3	`kissa kohtelu`	Treatment of cats	10	14	28
4	`ida aalbergin ura ulkomaa`	Ida Aalberg's career	38	93	90
16	`tulovero`	Income tax	8	9	14

We used a prototype (programmed with Python language) to expand the baseline query terms. The idea of this QE is to access an inflectional, noisy historical DCH collection. Our FCG was based on a static table of inflections produced for the words beforehand, but it could be performed dynamically. On top of FCG we performed the historical expansion and noise expansion. Next we will present the findings.

5 Findings

We used binary relevance with liberal relevance threshold (documents from levels R1, R2 and R3 are accepted as relevant) and retrieved the top-100 documents by Indri 5.0. The results are given in Table 7.

Table 7. Comparing the effectiveness of the short queries and automatically expanded queries. Legend: the abbreviation *cf* denotes the collection frequency of a term.

#	AP		P@10		QE terms produced			Range of *cf* for matching candidates
					search word to expand	number of expansion candidates existing in the index	number of expansion candidates not in index	
	basel.	QE	basel.	QE				
1	0.03	0.02	0.30	0.20	Australia alkuperäiskansa	13 0	70 58	1-328 -
2	0.03	0.16	0.20	0.50	diakonissalaitos	10	117	1-24
3	0.01	0.03	0.10	0.10	kissa kohtelu	21 16	37 42	1-1203 1-359
4	0.07	0.11	0.60	0.70	ida aalbergin ura ulkomaa	27 2 28 25	31 59 10 15	1-4263 1, 215 1-625 1-13896
16	0.00	0.14	0.00	0.40	tulovero	12	104	1-13

Table 7 makes it apparent that our method has some interesting potential. As explained in Section 4.3, we used only an extremely limited set of rules to produce expansion terms, but we were able to improve the search results in most cases. In particular, in case of topic 5 the baseline query *tulovero* (modern expression for the concept "*income tax*") gave a zero result. However, the top-5 query expansion words produced for this term (collection frequency of the word in parentheses) were as follows: *tuloweron (13), tuloweroa (12), tulowero (9), tulowerosta (4), tuloveron (4).* In this case, it is essential to observe that the top-4 query terms have an archaic spelling (with "w" instead of "v"). This suggests that combining FCG with noisy rules is a promising direction of development in DCH searching.

6 Discussion

End users often prefer expressing their information needs using few search words (Jansen et al., 2000). In noisy historical collections the words occur in many different surface forms. This is due to morphological variation, historical variation, and noise due to OCR errors. If historical collections are searched, it might be beneficial for the user to be able to utilize *several* such variants of the query terms – so that the query terms correspond to the level and type of variation present in the documents. (Järvelin et al., 2015) However, this is cumbersome to do intellectually and manually. Moreover, at the level of individual documents, various combinations of noise levels, types of historical variation, and morphological variation may be present. We have proposed utilizing *targeted query expansions* to simulate the "production" of surface level variation in different types of documents. Targeted QE entail three dimensions: inflectional, historical and OCR expansion dimensions. Figure 1 shows a simple prototypical design of a search interface for performing targeted QEs for user-given query terms.

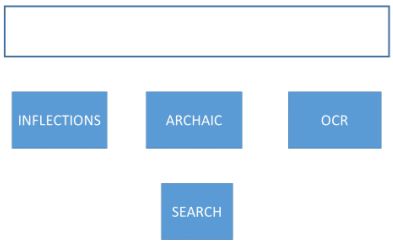

Figure 1. Design for the simple user interface for DCH search. The user enters the query terms into the search box. The subsequent user click (one of the blue buttons) will automatically produce different types of QEs. Each type of QE corresponds to a specific types of sub-collection.

We suggest that the searcher indicates the desired type of QE by pressing the corresponding interface button (and launching the search). The interface non-obtrusively supports performing QEs of desired types. For example, in the Indri search engine the set of expansion terms could be expressed as synonym-queries.

7 Conclusion

Large-scale production of digitized cultural heritage collections is on the way in many countries. The digitization process produces errors and therefore collections of mixed quality are created. To access these documents, query word expansion is needed, but manual and intellectual expansion is cumbersome for the end users. We have proposed a principle to perform focused query expansion towards digitized cultural heritage collections. We also suggest designing simple user interface to support the end user in expanding the queries towards different types of target documents.

References

M. Agosti, O. Conlan, N. Ferro, C. Hampson, G. Munnelly (2013a) Interacting with digital cultural heritage collections via annotations: the CULTURA approach. In Proc. of the 2013 ACM symposium on Document Engineering (DocEng '13), pp. 13-22, ACM Press, New York, NY, USA.

M. Agosti, M. Manfioletti, N. Orio, C. Ponchia, G. Silvello (2013b) The evaluation approach of IPSA@CULTURA. In Post-Proc. of the 9th Italian Research Conference (IRCDL 2013), Bridging Between Cultural Heritage Institutions Communications in Computer and Information Science, Revised Selected Papers, Vol. 385, 2014, pp. 147-152, Springer Berlin Heidelberg.

R. Alkula (2000) Merkkijonoista suomen kielen sanoiksi ("From strings to Finnish words"). Ph.D. Thesis, 295 pages, Acta Universitatis Tamperensis 763.
J. Callan, P. Kantor, D. Grossman (2002) Information retrieval and OCR: from converting content to grasping meaning. SIGIR forum, Vol. 36, No. 2, Fall 2002, pp. 58-61.
E. Di Buccio, G. M. Di Nunzio, G. Silvello (2014) A linked open data approach for geolinguistics applications, Int. J. Metadata, Semantics and Ontologies, Vol. 9, No. 1, pp. 29-41, 2014.
N. Ferro and G. Silvello (2014) Making it easier to discover, re-use and understand search engine experimental evaluation data, ERCIM News, Vol. 96, January 2014.
E. Hyvönen (2012) Publishing and using cultural heritage linked data on the semantic web, Morgan & Claypool Publishers, USA.
A. Järvelin, H. Keskustalo, E. Sormunen, K. Kettunen, M. Saastamoinen (2015) Information retrieval from historical newspaper collections in highly inflectional languages: a query expansion approach. Accepted for Journal of the Association for Information Science and Technology (JASIST).
P. Kantor, E. Voorhees (2000) The TREC-5 confusion track: comparing retrieval methods for scanned text. Information Retrieval, 2 (2/3), pp. 165-176.
H. Keskustalo, A. Pirkola, K. Visala, E. Leppänen, K. Järvelin (2003) Non-adjacent digrams improve matching of cross-lingual spelling variants. In: SPIRE 2003: 252-265.
K. Kettunen, E. Airio, K. Järvelin (2007) Restricted inflectional form generation in management of morphological keyword variation. Information Retrieval, 10 (4-5): pp. 415-444.
K. Kettunen, P. Arvola (2012) Generating variant keyword forms for a morphologically complex language leads to successful information retrieval with Finnish. IRFC 2012: 113-126.
K. Kettunen, T. Honkela, K. Lindén, P. Kauppinen, T. Pääkkönen, J. Kervinen (2014) Analyzing and improving the quality of a historical news collection using language technology and statistical machine learning methods. IFLA 2014, World Library and Information Congress. http://www.ifla.org/files/assets/newspapers/Geneva_2014/s6-honkela-en.pdf
A. Loponen, A. Pirkola, K. Järvelin, H. Keskustalo (2008) A novel implementation of the FITE-TRT translation method. In: ECIR 2008: 138-149.
M. Ohta, A. Takasu, J. Adachi (1997) Retrieval methods for english-text with missrecognized OCR characters. In: ICDAR'97: Proceedings of the 4th International Conference on Document Analysis and Recognition: 950-956.
A. Pirkola, H. Keskustalo, E. Leppänen, A-P. Känsälä, K. Järvelin (2002) Targeted s-gram matching: a novel n-gram matching technique for cross- and mono-lingual word form variants. Information Research, 7(2) (2002).
A. Pirkola, J. Toivonen, H. Keskustalo, K. Visala, K. Järvelin (2003) Fuzzy translation of cross-lingual spelling variants. In: SIGIR 2003: 345-352.
I. Raitanen (2012) "Etsikäät hywää ja älläät pahaa." Tiedonhakumenetelmien tuloksellisuuden vertailu merkki-virheitä sisältävässä historiallisessa sanomalehtikokoelmassa. ("Seek the good and not the bad." Comparison of effectiveness of information retrieval methods in a historical newspaper collection containing character errors.) M.Sc. Thesis, May 2012, University of Tampere, Finland. 68 pages.
J. Savoy, N. Naji (2011) Comparative information retrieval evaluation for scanned documents. In Proceedings of the 15th WSEAS international conference on Computers: 527-534.
X. Tong, C. Zhai, N. Milic-Frayling, D. Evans (1996) OCR correction and query expansion for retrieval on OCR data – CLARIT TREC-5 confusion track report. TREC 1996, 5 pages.