Aggregated Representation for the Focussed Retrieval of Structured Documents

Gabriella Kazai, Mounia Lalmas and Thomas Rölleke
Department of Computer Science
Queen Mary, University of London
London E1 4NS
{gabs, mounia, thor}@dcs.qmw.ac.uk

Abstract

Effective retrieval of structured documents should exploit the content and structural knowledge associated with the documents. This knowledge can be used to focus retrieval to the best entry points: document components that contain relevant information, and from which users can browse to retrieve further relevant components. To enable this, the representation of a document component is defined as the aggregation of the representation of its own content and the representation of its structurally related components. The aggregation makes it possible to specify how the representation of a component is influenced by that of its connected components. It also allows the capturing of the type of relationships between the components and the importance of components.

1 Introduction

With the widespread use of hypermedia and the rapid adoption of the XML markup language on the Web there is more scope and need to exploit the structural knowledge of documents for the purpose of their retrieval ([CMF96,Chi97,Dom01,FG01]). Numerous studies have highlighted that indexing web pages (e.g. [BH98,Sil00,YL96]) or structured documents in general (e.g. [Cal94,Wil94,Roe99,MJK+98]) based on combined structure and content knowledge can improve retrieval effectiveness. In addition, this combination makes it possible to retrieve relevant document components of varying granularity, for example, a document component when only that component is relevant, a group of components, when all the components in the group are relevant, or the document itself, when the entire document is relevant.

Structural knowledge can also be taken into account when displaying the retrieval results to the user. The retrieval results of an information retrieval (IR) engine in general are presented to the user as a ranked list of pointers. In traditional IR, the pointers provide links to whole documents (or web pages that are treated as whole documents) which contain relevant information to the user query. In structured document retrieval, the ranked list contains pointers to relevant document components. In both cases the retrieval results can contain objects that are "related" to each other, e.g. linked web documents or sub-components of the same document. According to the ranking method these related objects may be displayed at distant locations in the result. This can waste user time and lead to user disorientation [Chi97]. By exploiting structural knowledge these relationships can be made explicit to the user. One method is to list related objects in a sub-list (as followed by Google[1]). A second method is to group related objects into clusters (as performed by Northern Light Search[2]). A third method is to **focus retrieval** to so called **best entry points** [Lal97,CMF96]. These best entry points correspond to relevant document components from which users can browse to access further relevant document components Returning best entry points, and not merely relevant document components, is a means to capture relationships between retrieved document components.

In this paper, we develop an IR model that allows for the focussed retrieval of structured documents. The model provides a framework to formally capture and combine structural and content knowledge based on the notion of **aggregation** and criteria determining what constitutes a best entry point. We define the representation of a document component as the aggregated representation of its own content and the representation of its structurally related parts. The aggregation yields a model in which we can specify how the representation of a component is influenced by that of its connected components. It also allows the capturing of the types of relationships between components and the importance of components.

In a previous work [KLR01], we used an aggregation based on a fuzzy formalisation of linguistic quantifiers as proposed by [Yag00]. There the indexing criterion was expressed as follows: "index a document by a

[1] http://www.google.com/

[2] http://www.northernlight.com/

term if this term indexes Q of its related components”, where Q corresponds to a linguistic quantifier such as “most”, “all”, “at least one”. In the present paper, we use a more general aggregation formalism that provides a means to uniformly capture any evidence such as the importance of components, the types of links, etc in the aggregation. This leads to a more expressive formalism that supports the focussed retrieval of structured documents.

The remainder of the paper is organised into five sections. In Section 2 we provide an analysis of the basic concepts that influence the aggregation for focussed retrieval and form our indexing criterion. In Section 3 we define our aggregation formalism. We provide a framework to express the indexing criterion within the aggregation in Section 4. In Section 5, we draw parallels between the aggregation for focussed retrieval and the vector space model. Section 6 draws conclusions of our work and outlines future research.

2 Indexing Criterion

In this paper we consider a structured document as an acyclic directed graph of contexts, where a context is a document component of varying granularity including the whole document (root). A given context in the document graph can have a number of sub-contexts, where a sub-context is a document component that is connected to the composite context via an arc of the document graph. In an aggregation-based model for structured document retrieval, the representation of a composite context is defined as the aggregated representation of its own content and the representation of its sub-contexts. Our model must then provide means for representing the own content of a document context (i.e. the raw data contained in that component). It must then aggregate this representation with the aggregated representations of its sub-contexts. To support focussed retrieval the resulting representation of a document component should provide a measure of the component’s suitability as a best entry point. This requires criteria determining what constitutes a best entry point. We propose the following: the suitability of a context c as a best entry point with regards to a term t is a measure based on the term weights associated with the occurrences of the term t in c and in c’s sub-contexts, the importance associated with the sub-contexts, the types of link that connect c and its sub-contexts, and the portion and distribution of c’s sub-contexts that are indexed by the term t. This measure is calculated at indexing time and is expressed as the aggregated weight of an index term t in the context c. We then say that the indexing criterion is a function of the following factors:

1. The *term weight* associated with an index term t in a context c. We will denote this by $A_c(t)$. We will also refer to this weight as $A^*_{c0}(t)$, which defines the own content of the context c as the aggregated representation of a “special” sub-context of c. The value of this weight can be computed, for example, as the normalised term frequency of t in c.

2. The *aggregated term weights* associated with an index term t in the sub-contexts of a context c is given by $A^*_{c1}(t),\ldots,A^*_{cm(c)}(t)$ where $A^*_{ci}(t)$ is the aggregated term weight associated with t in the i-th sub-context of c, $c1..cm(c)$ denotes the sub-contexts of c and $m(c)$ is the number of sub-contexts of c. For example, consider a document d1 that consists of three sections, s1, s2 and s3, where the section s3 has two paragraphs, p1 and p2. When calculating the aggregated weight of t in d1, the aggregated term weights associated with t in the contexts s1, s2 and s3 are given by $A^*_{s1}(t)$, $A^*_{s2}(t)$ and $A^*_{s3}(t)$. By incorporating these measures into the indexing criterion we allow for our aggregation to take into account the term weights associated with the representation of a term t in c’s sub-contexts.

3. The importance associated with a context c, denoted by $P_e(c,c)$ or $P_e(c,c0)$, and the importance associated with c’s sub-contexts $c1..cm(c)$ with respect to c, denoted by $P_e(c,ci)$. The set of importance is then given as $I_c=\{P_e(c,c0),P_e(c,c1),P_e(c,c2),\ldots,P_e(c,cm(c))\}$. This measure is used to express which components are more important than others in contributing to the content of the composite component (e.g. an abstract might be more important than a discussion) and is common concern for structured document retrieval.

4. *Link type*, denoted by $P_l(l)$. The nature of the link (connection) between a context and its sub-contexts can also influence the aggregated weight. Links can be typed according to a number of characteristics and considerations. One could differentiate between structural and semantic links. Within the structural type we could further distinguish between hierarchical (e.g. sections and sub-sections) and linear links (next and previous sections). Link types could also be based on their popularity [BP98, Kl98]. A link type can then be associated with a weight or a weighting function, expressed through $P_l(l)$. This way, for example, we can give higher influence to hierarchically connected sub-contexts than linearly connected sub-contexts.

5. The *portion* and *distribution* of *c*'s sub-contexts indexed by the term *t*, denoted by $P_p(c,t)$ and $P_f(c,t)$, respectively. The idea here is to reflect in the aggregation what portion of a context's sub-contexts are indexed by the term *t* and how these sub-contexts are distributed. The higher number of sub-contexts are indexed by *t*, the higher the aggregated weight of *t* in the super-context. The distribution is important in order to differentiate between situations where the connected nodes indexed by *t* are uniformly distributed or when they form clusters. Say, for example, that a chapter has seven sections, out of which three are indexed with *t*. The aggregated weight should be higher if the three sections are evenly distributed (e.g. they are the second, fifth and seventh sections), and lower if the sections form a tight cluster (e.g. they are the first, second and third sections). When the connected components of a context *c* form clusters, then a cluster representative (one of the sub-contexts, a centroid or a metadocument etc.) is considered to be a better entry point than *c*. In summary, according to these two factors (portion and distribution), *c* would be considered a better entry point with regards to the term *t*, if a higher number of its sub-contexts are indexed by *t* and if these components are uniformly distributed[3].

The aggregation provides an implementation of the indexing criterion for representing structured documents for their focussed retrieval. According to this, a context in a structured document is represented as a collection of weighted index terms, where the weight of a term is calculated as the aggregated weight of all occurrences of that term in that context and in its sub-contexts. The indexing criterion upon which the aggregation is based makes it possible to obtain a model in which we can specify how the representation of a context is influenced by that of its sub-contexts. It allows the capturing of the type of links between the components (e.g. hierarchical vs linear), the importance of the components being aggregated (e.g. title vs. abstract), and the portion and distribution of the connected components indexed by a given term. With such a representation of structured documents, it then becomes possible to focus retrieval to the best entry points.

3 The Aggregation

The aggregation specifies the interrelationship between the aggregated arguments, the so-called *attributes*, and provides a mathematical way of combining their values to arrive at a single aggregated score.

<u>Definition I</u>: Let $S=(A_1,\ldots,A_m)$, $A_i \in [0,1]$ be the attribute space representing the arguments to be aggregated. The aggregation is defined by an *Aggregation Weighting Vector* $W=(w_1,\ldots,w_m)$, where $w_i \in [0,1]$ and $\sum_{i=1}^{m} w_i = 1$ such that

$$A^* = \sum_{i=1}^{m} w_i \cdot A_i$$

where A^* is the aggregated score, A_i is the *i*-th attribute in *S*, and w_i is the *i*-th weight in *W*.

In the context of structured document retrieval, and based on this definition, we define the aggregated weight of a term *t* in a context *c*, as the weighted sum of the attributes, $A^*_{c0}(t), A^*_{c1}(t),\ldots,A^*_{cm(c)}(t)$, where $A^*_{c0}(t)$ is the term weight of *t* in the own content of *c* and $A^*_{ci}(t)$ is the aggregated term weight of *t* in *i*-th the sub-context of *c*. This is formally captured in the following definition.

<u>Definition II</u>: Let $S_c(t)=(A^*_{c0}(t),A^*_{c1}(t),\ldots,A^*_{cm(c)}(t))$, $A^*_{ci}(t) \in [0,1]$ be the attribute space representing the arguments to be aggregated, where $A^*_{c0}(t)$ is the term weight of *t* in *c*'s own content and $A^*_{ci}(t)$ (for *i*=1 to *m*(*c*)) is the aggregated weight of *t* in the *i*-th sub-context of *c*, and *m*(*c*) is the number of sub-contexts of *c*. The aggregation is defined by an *Aggregation Weighting Vector* $W_c(t)=(w_{c0}(t),w_{c1}(t),\ldots,w_{cm(c)}(t))$, where $w_{ci}(t) \in [0,1]$ and $\sum_{i=0}^{m(c)} w_{ci}(t) = 1$ such that

$$A^*_c(t) = \sum_{i=0}^{m(c)} w_{ci}(t) \cdot A^*_{ci}(t)$$

where $A^*_c(t)$ is the aggregated term weight of the term *t* in *c*, and $w_{ci}(t)$ is the *i*-th weight in $W_c(t)$. Using vector notation this can be expressed as the scalar product of $W_c(t)$ and $S_c(t)$: $A^*_c(t) = \overrightarrow{W_c(t)} \cdot \overrightarrow{S_c(t)}$.

This means that we associate a context, *c*, with a set of (attribute, term) pairs, $\{(A_c(t_i), t_i)\}$, where $A_c(t_i)$ corresponds to term weight of t_i in *c* and reflects the extent to which the term t_i is a good content descriptor

[3] User studies are necessary to validate these views.

of c and is calculated by some estimated measure of relevance. The aggregated representation of c is given by the set of (attribute, term) pairs, $\{(A^*_c(t_i), t_i)\}$, where $A^*_c(t_i)$ refers to the aggregated term weight of the term t_i in c which reflects the extent to which the context c is a best entry point with regards to the term t_i, and is calculated as the aggregation of the term weights associated with the occurrences of t_i in c and in c's sub-contexts. In order to derive the full representation of a document context, c, we need to apply the aggregation to all terms t that index c or any of its sub-contexts. The weight of all other terms in the term space will be assigned a weight of 0 in c.

Next we describe how the parameters of the indexing criterion can be expressed within the aggregation framework. By providing an algorithmic way to obtain the aggregated weight of a term in a document component, we can provide a uniform implementation of aggregation based on the indexing criterion.

4 Expressing the Indexing Criterion in the Aggregation

As mentioned in Section 2 the indexing criterion is a function of the following aggregation parameters: $A^*_{c0}(t)$: the term weight associated with a term t representing the own content of a context c; $A^*_{c1}(t),\ldots,A^*_{cm(c)}(t)$: the aggregated term weights associated with the occurrences of t in c's sub-contexts; $I_c=(P_e(c,c0),P_e(c,c1),\ldots,P_e(c,cm(c)))$: the component's importance with respect to the context c; $P_l(l)$: the link types connecting the sub-contexts with c; and $P_p(c,t)$ and $P_f(c,t)$: the portion and distribution of c's sub-contexts indexed by t. The aggregation must implement all of these elements of the indexing criterion. As defined in Section 3, the term weights associated with the document context's own content and its sub-contexts' representation are already expressed as the elements of attribute vector, $S_c(t)$. The remaining parameters of the indexing criterion will be expressed within the aggregation weighting vector, $W_c(t)$. We will now first look at some potential methods for calculating the individual parameters then show one possible method for deriving the aggregation weighting vector based on their combined influence.

4.1 Estimating the aggregation parameters $P_p(c,t)$, $P_f(c,t)$, $P_l(l)$ and $P_e(c,ci)$

4.1.1 *$P_l(l)$: link types*

As we mentioned earlier link types can be distinguished based on a number of considerations such as the popularity of the link. Although the categorisation of link types is subject to on-going research, there are two commonly distinguished link types: structural and semantic. Structural links act as navigational links and connect documents or document components within the same domain. They can be further categorised into composition (hierarchical) or sequence (linear) types [GC01]. Semantic links connect documents with similar topics. The general approach is to conclude that when two linked document contexts belong to different domains, then the link between them is a semantic link. In [KLR01] we distinguished three types of links: hierarchical, linear and referential. However, other criteria could also play an important part in categorising links. For example, links in a HTML document can be typed according to the formatting of the anchor text (bold, italics, large font size etc.) and the positioning of the link on the page.

We aim to assign different weights to the different link types to control how much the linked context should contribute to the aggregation. This way, for example, a web page pointed to by a hardly noticeable link could be considered less important than a web page linked by a prominently placed and formatted link.

At the moment we have no concrete evidence on what characteristics of links should be used in establishing link types and what weighting schema should be applied. Further research would be needed to establish a general model for categorising link types and to arrive at an appropriate weighting model. To support focussed retrieval and the selection of best entry points research in users' browsing behaviour is also necessary.

4.1.2 *$P_e(c,ci)$: context importance*

This parameter reflects the importance of a sub-context of a context c with respect to c. The aim of the parameter is to allow for the different sub-contexts of c to contribute with different emphasis to the representation of c. For example, we could assign higher importance weights to titles, abstracts and bibliographical sections if document representations based on them were shown to improve retrieval effectiveness. When considering the representation of smaller contexts, like paragraphs, based on [Bax58], we could assign higher importance weights to sentences located at the beginning of a paragraph, as these are believed to contain the central theme of that paragraph.

The problem of deriving a weight to assign to a context in order to reflect its importance is, however, an unsolved issue and ground for further research. The problem is twofold. Firstly, it requires knowledge on which sub-contexts are more important than others (e.g. is an abstract really more important than a conclusion or is a title more important than the full text etc.). Furthermore it requires a method for quantifying this importance by assigning a value between 0 and 1, where 0 represents that the context is not at all important and 1 means that it is very important. Values between 0 and 1 would reflect that the context is somewhat important.

4.1.3 $P_p(c,t)$: the portion parameter

We made the assumption that the aggregated weight of a term *t* in a context *c* should be high if a high portion of *c*'s sub-contexts is indexed by *t*. This seems intuitive and is also supported by the preliminary findings of a user study we conducted as part of the Focus project (http://qmir.dcs.qmw.ac.uk/Focus/index.htm).

A number of functions can be employed to provide a measure for this parameter. We could simply calculate the ratio of the sub-contexts indexed by *t* and the total number of sub-contexts. Another method would be to take the lowest attribute value, $Min(S_c(t))$. This would mean that a context would only be indexed by the term *t* if "all" of its sub-contexts were indexed by *t*. An alternative is to use the average function, $Avg(S_c(t))$. This would result in a higher weight if more of *c*'s connected components are indexed by *t*. In [KLR01] we used fuzzy representations of linguistic quantifiers as a means to calculate this parameter.

4.1.4 $P_f(c,t)$: the distribution parameter

A measure of distribution is only required if a connection (e.g. linear link) exists between the sub-contexts of a context *c*, as is most often the case with structured documents, where one section follows another. The list of web documents pointed to by links on a web page can also be regarded linearly linked if we consider the order of the links on the page.

When measuring the distribution of sub-contexts indexed by the term *t* among the whole set of sub-contexts, we aim to arrive at a function that will result in higher weights if the distribution is uniform and lower weights if the components form dense clusters. We can say that the distribution is uniform if there are no "big gaps" between components indexed by *t*. A gap in this sense is a set of document components that are not indexed by *t* and are positioned between two document components indexed by *t* within the linearly connected chain. The size of a gap is given by the number of sub-contexts that form a contiguous section. We can then base the measure of distribution on the following formula

$$P_f(c,t) = \frac{\sum_{i=1}^{m(c)-1} f_{i+1}(c,t) - f_i(c,t)}{m(c)-1} \quad , m(c) > 1$$

where $P_f(c,t)$ is a measure of distribution given a term *t* defined upon the sub-contexts of a context *c*, $f_i(c,t)$ is 1 if *t* occurs in the *i*-th sub-context of *c* and 0 otherwise. Take the following as an example. Let us say that a chapter of a document has twelve linearly connected sections, where five of the sections are indexed by *t*. The following are two possible distributions.

Scenario 1:	**s1**	s2	s3	**s4**	**s5**	s6	s7	**s8**	s9	**s10**	s11	s12
Scenario 2:	**s1**	**s2**	**s3**	**s4**	s5	s6	s7	s8	s9	s10	**s11**	s12

In scenario 1 the distribution is $P_f(c,t)$=(1+0+1+0+1+0+1+1+1+1+0)/11=0.63, whereas in scenario 2 we obtain $P_f(c,t)$=(0+0+0+1+0+0+0+0+0+1+1)/11=0.27. This reflects that the sections in scenario 1 are more evenly distributed than in scenario 2. Of course, other methods of measuring the distribution can also be implemented.

4.2 Combining the aggregation parameters $P_p(c,t)$, $P_f(c,t)$, $P_l(l)$ and $P_e(c,ci)$

In this paper, we identified four sources of evidence for estimating the values of the $W_c(t)$ vector (see Definition II). With each source, we associate a measure as follows, where *Context*, *Term* and *LinkType* correspond to the set of contexts, terms and link types, respectively:

- $P_e(c,ci)$: defines the importance of a sub-context with regards to a context and is a mapping of *Context* × *Context* → [0,1]
- $P_l(l)$: reflects the overall probability (importance) of a link type (structural, referential, etc) and is a mapping of *LinkType* → [0,1]
- $P_p(c,t)$: provides a measure for the relative number of sub-contexts of a context in which a term occurs and is a mapping of *Context* × *Term* → [0,1]
- $P_f(c,t)$: measures the distribution (frequency) of a term among the sub-contexts of a context and is a mapping of *Context* × *Term* → [0,1]

We recall that the aggregated weight of a term is estimated by the scalar product of the aggregation weighting vector $W_c(t)$ and the attribute vector $S_c(t)$, where each of the vector components corresponds to a value for a term in a sub-context (*Context* × *Term*). With respect to our aggregation framework, we use the aforementioned sources of evidence for estimating the aggregation vector $W_c(t)$.

Given a term t and a context c, the aggregation vector $W_c(t)$ contains a component for each sub-context of c and one for c itself. The combined evidence of the parameters $P_e(c,ci)$, $P_l(l)$, $P_p(c,t)$, $P_f(c,t)$ shall produce the values of $W_c(t)$. A very simplistic approach would be the normalised linear combination (weighted sum) of the parameters.

$$w_c(t) = \frac{v_1 \cdot P_e(c,ci) + v_2 \cdot P_l(l) + v_3 \cdot P_p(c,t) + v_4 \cdot P_f(c,t)}{\sum_{m(c)+1} v_1 \cdot P_e(c,ci) + v_2 \cdot P_l(l) + v_3 \cdot P_p(c,t) + v_4 \cdot P_f(c,t)}$$

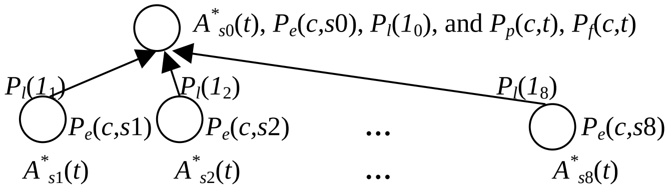
Of course, it remains the task of determining the values v_i of the linear combination. This could be assigned experimentally or could be based on a learning function. As the main goal of this paper was to describe an aggregation framework, the main goal of our experimental work will be to find an estimate of the $w_c(t)$ values (see for example the methodology proposed in [RLQ01]).

4.3 Aggregating - an example

In this section we use a simple example to demonstrate how the aggregation operator can be applied to derive the representation of a document component that supports focussed retrieval. Consider the document component, c, which has eight sections. The term weight associated with t in the own content of c is $A_c(t)$, which is represented by $A^*_{s0}(t)$. The aggregated term weights associated with the representations of the term t in the sections are given below along with $A^*_{s0}(t)$.

$A^*_{s0}(t)$	$A^*_{s1}(t)$	$A^*_{s2}(t)$	$A^*_{s3}(t)$	$A^*_{s4}(t)$	$A^*_{s5}(t)$	$A^*_{s6}(t)$	$A^*_{s7}(t)$	$A^*_{s8}(t)$
0.4	0.9	0	0.1	0	0	0.9	0.5	0.8

In order to derive the aggregated representation, $A^*_c(t)$, of the term t in the context c, we need to aggregate the above representations of t taking into account the parameters: $P_e(c,ci)$, $P_l(l)$, $P_p(c,t)$ and $P_f(c,t)$, which form the aggregation weighting vector. The attributes and parameters of the aggregation are summarised in the diagram below.



For simplicity we assume that each link is of the same type (hierarchical) and has a weight of 1, (i.e. $P_l(l_0)$= $P_l(l_1)$=...=$P_l(l_8)$=1). The importance associated with each of the components is:

$P_e(c,s0)$	$P_e(c,s1)$	$P_e(c,s2)$	$P_e(c,s3)$	$P_e(c,s4)$	$P_e(c,s5)$	$P_e(c,s6)$	$P_e(c,s7)$	$P_e(c,s8)$
1	0.9	0.9	0.3	0.2	0.1	0.1	0.4	0.8

We will use the $Avg(S_c(t))$ function to measure the portion of connected components that are indexed by t. This results in $P_p(c,t)$=(0.4+0.9+0+0.1+0+0+0.9+0.5+0.8)/9=0.4. The distribution is calculated according to the formula given in Section 4.1.4 and is equal to $P_d(c,t)$=(0+1+1+1+0+1+0+0)/8=0.5. In this example, we use the weighted sum of Section 4.2 to combine the parameters in order to derive the elements of the aggregation weighting vector $W_c(t)=(w_1,..,w_9)$. We assign $v_1=v_2=v_3=v_4=0.25$ to allow for all four parameters to equally contribute to $W_c(t)$. The aggregation weighting vector is then given as:

w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9
(0.25·0.4+ 0.25·0.5+ 0.25·1+ 0.25·1) /Σ $W_c(t)$ =0.134	0.25·0.4+ 0.25·0.5+ 0.25·1+ 0.25·0.9 /Σ $W_c(t)$ =0.128	0.25·0.4+ 0.25·0.5+ 0.25·1+ 0.25·0.9 /Σ $W_c(t)$ =0.128	0.25·0.4+ 0.25·0.5+ 0.25·1+ 0.25·0.3 /Σ $W_c(t)$ =0.101	0.25·0.4+ 0.25·0.5+ 0.25·1+ 0.25·0.2 /Σ $W_c(t)$ =0.096	0.25·0.4+ 0.25·0.5+ 0.25·1+ 0.25·0.1 /Σ $W_c(t)$ =0.091	0.25·0.4+ 0.25·0.5+ 0.25·1+ 0.25·0.1 /Σ $W_c(t)$ =0.091	0.25·0.4+ 0.25·0.5+ 0.25·1+ 0.25·0.4 /Σ $W_c(t)$ =0.107	0.25·0.4+ 0.25·0.5+ 0.25·1+ 0.25·0.8 /Σ $W_c(t)$ =0.124

The aggregated weight for t in c is then equal to $A^*_c(t)$=0.4·0.134+0.9·0.128+0·0.128+0.1·0.101+0·0.096+ 0·0.091+0.9·0.091+0.5·0.107+0.9·0.124=0.4135. This reflects that the document component c has an associated measure of 0.4135 to be a suitable best entry point with regards to the term t. This relatively low score means that c is a less suitable best entry point with regards the term t into the above document structure. The reason for this is the clustered nature of the sub-contexts and the low component importance values (see s6 for example).

5 Parallels with the vector space model

In this section, we highlight parallels between our model and the vector space model. In the latter, the relevance of documents to a query is based upon the scalar product of the document-term matrix and the query vector (RSV stands for relevance status value):

$$\overrightarrow{RSV(D,q)} = D \cdot \vec{q} = \begin{bmatrix} \overrightarrow{d_1} \\ \overrightarrow{d_2} \\ \dots \\ \overrightarrow{d_n} \end{bmatrix} \cdot \vec{q} = \begin{bmatrix} w_1(t_1) & w_1(t_2) & \dots & w_1(t_p) \\ w_2(t_1) & w_2(t_2) & \dots & w_2(t_p) \\ \dots & \dots & \dots & \dots \\ w_n(t_1) & w_n(t_2) & \dots & w_n(t_p) \end{bmatrix} \cdot \vec{q}$$

where D is the document-term matrix and the d_i are p dimensional document vectors, where p is the number of distinct terms in the document collection.:

In our aggregation framework, we consider a matrix S_c of vectors $S_c(t_i)$ of aggregated term weights for each context c and term t_i. The scalar product of an aggregated term vector $S_c(t_i)$ with the aggregation vector $W_c(t_i)$ yields the aggregated weight of a term $A^*_c(t_i)$. In vector notation $A^*_c(t_i) = \overrightarrow{W_c(t_i)} \cdot \overrightarrow{S_c(t_i)}$, that is:

$$\overrightarrow{A^*_c} = \begin{bmatrix} A^*_c(t_1) \\ A^*_c(t_2) \\ \dots \\ A^*_c(t_p) \end{bmatrix} = \begin{bmatrix} \overrightarrow{S_c(t_1)} \\ \overrightarrow{S_c(t_2)} \\ \dots \\ \overrightarrow{S_c(t_p)} \end{bmatrix} \cdot \vec{W_c} = \begin{bmatrix} A^*_{c0}(t_1) & A^*_{c1}(t_1) & \dots & A^*_{cm(c)}(t_1) \\ A^*_{c0}(t_2) & A^*_{c1}(t_2) & \dots & A^*_{cm(c)}(t_2) \\ \dots & \dots & \dots & \dots \\ A^*_{c0}(t_p) & A^*_{c1}(t_p) & \dots & A^*_{cm(c)}(t_p) \end{bmatrix} \cdot \vec{W_c}$$

where $S_c(t_i)$ corresponds to the vector $(A^*_{c0}(t_i),A^*_{c1}(t_i),\ldots,A^*_{cm(c)}(t_i))$, $m(c)$ being the number of related contexts of c. The aggregation vector W_c corresponds to the query vector in the vector space model. The matrix S corresponds to the document-term matrix D of the vector space model.

With respect to the focussed retrieval of structured documents, the aim is to find the best aggregation vector so to obtain the “best aggregated term weights” to determine the best entry points. With respect to classical document retrieval, the aim is to find the best query formulation (e.g. query term weighting, query term expansion) so to obtain the best relevance weights to determine the relevant documents.

Now consider all contexts in a collection of structured documents, $c_{d1}...c_{dn}$, where n is the total number of contexts. We obtain the document-term matrix D from the vectors $\overrightarrow{A^*_{cdi}}$ of aggregated term weights for each context c_{di}.

$$D = \left[\overrightarrow{A^*_{Cd1}} \; \overrightarrow{A^*_{Cd2}} \ldots \overrightarrow{A^*_{Cdn}} \right]$$

The document-term matrix contains a term weight vector (row) for each context occurring in the collection. With this formalisation we have established a direct link between the vector-space model and structured document retrieval based on aggregated term weights.

6 Conclusion and Future Work

This paper describes a model for the representation of structured documents to allow for their focussed retrieval, that is the identification of best entry points. Particular attention is paid to the representation through the formalisation of an indexing criterion that takes into account the types of links between document components, the importance of components forming the structured documents, and a measure of proportion and distribution of terms in related components. The formalisation is based on an aggregation framework that allows combining in a uniform fashion all evidence that may be used to determine what constitutes a best entry point.

We will implement our approach to evaluate its effectiveness in retrieving best entry points. We are currently eliciting criteria based on user studies that determine what is a best entry point in a structured document for a given query (http://qmir.dcs.qmw.ac.uk/Focus/index.htm). Experiments will be carried out on a test collection of XML documents that we built. This test collection consists of 37 Shakespeare plays marked up in XML by Jon Bosak[4]. We have 43 queries addressing English and Drama students' real information needs. We have obtained relevance assessments for these queries, and are currently gathering user criteria for focussed retrieval. These will provide us with best entry points for the 43 queries, as well as insights regarding the selection of the aggregation parameters.

Acknowledgement

This work has been carried out in the framework of the EPSRC GR/N37612 and British Council ARC Project 1162.

References

[Bax58] P B Baxendale. Man-made index for Technical Literature - an Experiment. *IBM Journal of Research and Development, Vol 2(4)*, pp 354-361. 1958.

[BH98] K Bharat and M Henzinger. Improved Algorithms for Topic Distillation in Hyperlinked Environments, *Proceedings of the International ACM-SIGIR Conference*, 1998.

[BP98] Brin, S. & Page, L. The Anatomy of a Large-Scale HyperTextual Web Search Engine. *Proceedings of the Seventh International World Wide Web Conference*, Brisbane, Australia, 1998.

[Cal94] J Callan. Passage-Level Evidence in Document Retrieval. *ACM SIGIR*, pp 302-310, Dublin, 1994.

[Chi97] Y Chiaramella. Browsing and Querying: Two Complementary Approaches for Multimedia Information Retrieval. *Hypermedia - Information Retrieval – Multimedia*, Dortmund, Germany, 1997.

[CMF96] Y Chiaramella, P Mulhem and F Fourel. A Model for Multimedia Information Retrieval. Technical Report Fermi ESPRIT BRA 8134, University of Glasgow, 1996.

[Dom01] S Dominich. *Mathematical Foundation of Information Retrieval*. Kluwer Academic Press, 2001.

[FG01] N Fuhr and K Großjohann. XIRQL: A Query Language for Information Retrieval in XML Documents. *ACM-SIGIR*, New Orleans, 2001

[Kl98] Kleinberg, J. Authoritative sources in a hyperlinked environment. *Proc. 9th ACM-SIAM Symposium on Discrete Algorithms*, 1998. Extended version in *Journal of the ACM* 46, 1999.

[KLR01] G Kazai, M Lalmas and T Roelleke. A Model for the Representation and Focussed Retrieval of Structured Documents based on Fuzzy Aggregation. *Proceedings of SPIRE*, 2001, Chile

[4] http://www.ibiblio.org/bosak/

(To appear).

[Lal97] M Lalmas. Dempster-Shafer's Theory of Evidence Applied to Structured Documents: Modelling Uncertainty, *Proceedings of ACM SIGIR*, pp. 110-118, Philadelphia, 1997.

[MJK+98] S Myaeng, DH Jang, MS Kim and ZC Zhoo. A Flexible Model for Retrieval of SGML Documents. *ACM-SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, pp 138-145, 1998.

[RLQ01] T Roelleke, M Lalmas and S Quicker. The Accessibility Dimension in Structured Document Retrieval, 2001 (Submitted for publication).

[Roe99] T Roelleke. *POOL: Probabilistic Object-Oriented Logical Representation and Retrieval of Complex Objects - A Model for Hypermedia Retrieval*, Ph.D. Thesis, University of Dortmund, Verlag-Shaker, 1999.

[Sil00] I Silva, B Ribeiro-Neto, P Calado, E Moura and N Ziviani. Link-Based and Content-Based Evidential Information in a Belief Network Model. *Proceedings of ACM-SIGIR 23rd*, Athens, 2000.

[Wil94] R Wilkinson. Effective Retrieval of Structured Documents. *ACM-SIGIR*, Dublin, pp 311-317, 1994.

[Yag00] RR Yager. A Framework for Linguistic and Hierarchical Queries in Document Retrieval. *In Soft Computing in Information Retrieval – Techniques and Applications*, pp3-20, 2000.

[YL96] B Yuwono and D. L. Lee. Search and Ranking Algorithms for Locating Resources on the World Wide Web. *Proceedings of the International Conference on Data Engineering (ICDE),* New Orleans, 1996.