The Open University

Open Research Online

The Open University's repository of research publications
and other research outputs

Exploring pathways across stories

Conference or Workshop Item

How to cite:

Zdrahal, Zdenek; Mulholland, Paul and Collins, Trevor (2008). Exploring pathways across stories. In: IEEE 2008 International Conference on Distributed Human-Machine Systems, 10-12 Mar 2008, Athens, Greece.

For guidance on citations see FAQs.

Version: Version of Record

Link(s) to article on publisher's website:
http://kmi.open.ac.uk/people/paulm/docs/zdrahal-DHMS08.pdf

oro.open.ac.uk

Exploring Pathways Across Stories

Zdenek Zdrahal, Paul Mulholland and Trevor Collins
Knowledge Media Institute, The Open University, UK
{Z.Zdrahal, P.Mulholland, T.D.Collins}@open.ac.uk

Abstract— **This paper describes a method for supporting the exploration of a collection of documents organized as a hypertext by investigating relations between documents along user-specified paths. The approach is demonstrated on a corpus of stories about the World War Two activities of the British Government Code and Cypher School at Bletchley Park. Each story is described by one or more events and annotated in terms of domain ontologies. A pathway in the document space is a sequence of events in which adjacent events share common binding concepts. The criteria for selecting the pathway include a measure of the adherence to the user-specified part of the document space and the mutual information between adjacent documents calculated from their annotations.**

Index Terms—**pathway, document space, ontology, mutual information, seed term, focused document space.**

I. INTRODUCTION

The exploration of a large collection of documents is a task of increasing practical importance. Due to the low effort of electronic publishing there is a good chance that any topic of interest has been described somewhere in a digital library, document database or on the web. However, it is unlikely that the information we are looking for is contained in a single document. We may rather expect that it will be scattered across many resources with different styles, media and authors. In order to get an answer to our questions, search by means of keywords must be complemented by the exploration of a document set. There are other interpretations of the same problem. For example, an event which happened in the past has been independently described by many individuals who played different roles and represented different viewpoints. The goal is not to answer a single question but to reconstruct the past from such fragmented information and comprehend what happened. Examples may include the interpretation of historical events (see our case study in section II), an analysis of police protocols, and similar type of problems.

We call the collection of all documents that are available for exploration the *document space*. In this paper, the document space is a collection of stories. The term "story" refers to a semantically self-contained block of text, possibly with associate pictures or multimedia. There are two main reasons for concentrating on stories: knowledge is often represented in

stories [11], and there is a natural way of breaking the document into smaller units - events. However, a similar approach can be used for documents without the story structure [10].

We assume that the documents in the collection are semantically interrelated i.e. they share common key concepts and refer to related events. The whole document space can be regarded as a form of hypertext with the stories as lexias [8], and links defined by common concepts. Unlike in standard hypertext, in our method the links between lexias are not explicitly predefined, but are extracted from annotations of documents and constructed dynamically in accordance with the user's needs.

In addition to the knowledge extracted from individual documents, further meaning can be inferred from organising and presenting a set of documents in different structures. Imagine three stories describing events in a company: the first story is about someone's job interview, the second one, about the first days in a new job and the third one, about work experience a few years later. Organising these stories chronologically into a path allows the user to reason about the potential of personal development in the company, while comparing and contrasting second and third stories highlights the difference between the attitude of novices and experienced workers. Note, that the structure does not add any new knowledge. It only facilitates the discovery of the knowledge already present in the collection.

Our approach is based on three assumptions:
- Each document has its own meaning and the user can learn something new by reading it.
- There are inter-document dependencies (as described above) that make it possible to extract additional meaning from the set.
- Documents are annotated in terms of an appropriate domain ontology. This means that we do not analyse text by natural language processing techniques, but our inferences are based on the document annotation.

Two techniques often used for content exploration are clustering of content and creating pathways in the document space. In the CACM special issue on exploratory search they are represented by articles [4] and [6], respectively. Our approach to clustering documents (stories), followed by constructing semantic categories and intra-cluster hierarchies has been described in [10]. In this paper we present methods for content exploration by analyzing linear paths in the document space.

The rest of the paper is organized as follows. In Section 2 we describe the application domain. Section 3 introduces a formal representation of the document space as a hypergraph and exploratory pathways as paths in the hypergraph. Reducing the size of the document space by 'focusing' is presented in Section 4. Mutual information as a criterion for assessing the relationship between documents along the pathway is proposed in Section 5. Related approaches are described in Section 6 and in Section 7 we present the conclusions.

II. Case Study: Stories about Cultural Heritage

Constructing exploratory pathways from a collection of stories will be demonstrated on the case study developed for Bletchley Park as a part of the cultural heritage project CIPHER (FP5 IST-2001-32559). The documents relate to the history of Bletchley Park

A. The Bletchley Park Story

During World War Two, Bletchley Park was the location of the Headquarters of the British Government Code and Cypher School and hosted a number of distinguished scientists who worked on breaking enemy codes. In the early 1990s, the place was converted into a museum. Nowadays, the visitors are taken on a guided tour through the Bletchley Park Mansion and Huts in which people used to work. As a part of their research, the guides also collect relevant historical documents. These include transcripts of interviews with former Bletchley Park staff and wartime documents that have been recently released to the National Archive. At present, the Bletchley Park collection consists of thousands of unique documents about code breaking, life and work of prominent scientists and ordinary staff in Bletchley, impact of the Bletchley Park effort on the course of the war in different parts of the world and other associated topics.

B. Bletchley Park Text

Content exploration by analyzing pathways has been applied to the set of stories collected by the tour guides and made available to the museum visitors. The application developed for the Bletchley Park museum is called *Bletchley Park Text*. Out of the total of a few thousand stories about 400 of the most interesting stories were selected. They were annotated in terms of domain ontologies, with the CIDOC Conceptual Reference Model (CRM) as the upper ontology [2]. The next level ontologies include the *bletchley-park-ontology* which specialize the CRM concepts for the domain of the Bletchley Park Museum and the *narrative-ontology* which describes concepts used for linking stories, story objects and their media presentation. The selected stories were further divided into events as the basic knowledge units and annotated in terms of concepts, such as actors, places, objects, roles and time. Events are meaningful parts of the story describing an activity. A typical story consists of two or three events. An example event annotation is shown in figure 1.

Unlike a standard hypertext, in Bletchley Park Text the lexia are not connected by static links. Documents are linked dynamically - links are created only when the structure is needed. Links are defined in terms of *binding concepts* which are instances of selected ontology classes. At present, we use instances of classes *actor*, *place* and *object* as the binding concepts.

A pathway is a sequence of stories in which two adjacent stories share a binding concept. Though the annotation of the Bletchley Park content in terms of the CRM ontology provides a rich description of domain objects, the binding concepts are only a subset of the available CRM descriptors.

instance-of	bletchley-park-life-experience
name	jc-event-1
has-actor	peter-hilton donald-michie ralph-tester
has-billeting-location	bedford
has-recipient	double-playfair-code
has-time-specification	1942-1943
has-work-location	bletchley-park

Fig. 1. Event annotation

Slot names and types do not participate in creating links, but are used to annotate links and explain the role of binding objects in the pathway.

III. Formal representation

The document space can be represented as a hypergraph $H = <C,E>$, where $C = \{c_1, ... , c_N\}$ is a set of nodes corresponding to concepts and $E = \{e_1, ..., e_M\}$ is a set of edges corresponding to events (stories, documents) [1]. Figure 2 shows a hypergraph with 13 concepts and 5 annotated documents.
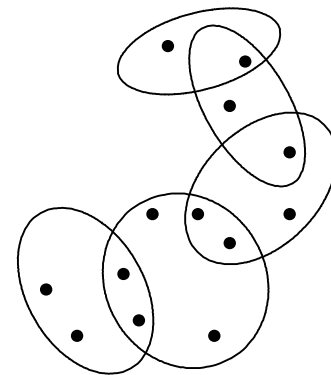

Fig. 2. Representing document space by a hypergraph

The document hypergraph is constructed as follows:
1. Annotated events in all documents specify the set of edges, $E = \{e_1, ..., e_M\}$. They are represented as n-tuples of extracted binding concepts (actors, objects and places) with associated event names for easy identification.

2. The set of nodes $C = \{c_1, \ldots, c_N\}$ is defined as a union of all edges e_i, i.e. $C = \cup e_i$. The slot names of event classes are ignored.

A *path* of length q in a hypergraph is a sequence of q different nodes of C so that any two adjacent nodes are members of the same edge in E. If there is a path between any two nodes the hypergraph is said to be *connected*. A similar statement can be made about the document space. Pathways in the documents space correspond to paths in the graph. We have analysed the Bletchley Park document space and experimentally verified that it is *connected* and consists of 1660 concepts and 770 events. The shortest pathway between any two concepts has length lower or equal to 7.

Content could be further divided into a number of themes or topics. In the document space themes correspond to clusters of documents calculated in accordance with some suitable intra/inter cluster measure of distance. In the Bletchley Park story collection themes include code breaking, wartime life in England, early computing, war in Africa, Pacific and a few others. In the hypergraph representation, the existence of a theme is manifested by a non-uniform distribution of overlapping edges associated with a group of concepts.

IV. EXPLORING PATHWAYS

Even if we restrict the way of binding documents as described above, there are many strategies of calculating pathways in the document space. Depending on the selected criterion, pathways play different roles in content exploration.

A. Simple path between two concepts

Conceptually the simplest structure is the linear pathway connecting two documents in the document space without any further restriction. It is a sequence $< c_1, e_1, c_2, e_2, \ldots, e_{R-1}, c_R >$ where c_1 is the initial binding concept and c_R is the terminal binding concept and for any two adjacent concepts $c_i, c_j \in e_k$ for some k.. Since the Bletchley Park hypergraph is connected, a pathway between any two concepts in the document space always exists. Potentially, there can be many pathways between the same two concepts. In the simplest version we provide the user with the shortest pathway between two selected concepts.

B. Focusing document space

The document space typically consists of many interrelated themes. The users are often interested only in some of them but would like to explore them in detail. The choice can be made by marking a few *seed concepts* which implicitly characterise the themes of interest. The seed concepts could be any binding concepts used to restrict the document space and focus the exploration.

Based on a selected set of seed concepts the document space is reconstructed by the following algorithm:

Let $S = \{c_1, c_2, \ldots, c_s\}$ be a set of seed concepts selected from the set of nodes C. Let us denote as $E_S = \{e_1, \ldots, e_S\}$ the set of all edges that contain at least one concept of S and C_S

the set of all nodes of E_S, $C_S = \cup e_i$ for all $e_i \in E_S$. Then hypergraph $H_S = < C_S, E_S >$ is a *partial hypergraph* of H [1] and the corresponding subspace of the original document space is called *focused document space*. An example of seed concepts and the corresponding partial hypergraph is shown in figure 3.

If all seed concepts characterise the same theme, then the focused document space contains only concepts of the corresponding cluster and paths in H_S might be constructed within this theme. However, as the clusters (themes) are overlapping seed concepts may select multiple themes.
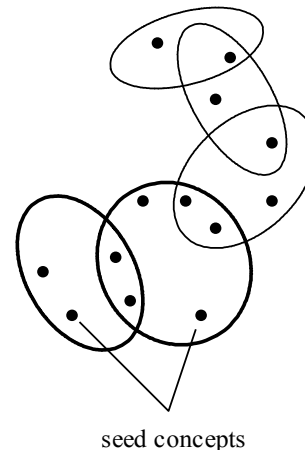

seed concepts

Fig. 3. Focusing: seed concepts and the partial hypergraph

Focusing reduces the document space to be explored. If seed concepts hit two themes described by disjoined clusters, the originally connected document space becomes disconnected and there are concepts in the focused document space for which a pathway does not exist.

C. Focusing document space in Bletchley Park Text

In Bletchley Park, the tour guides present the visitors with a history of the place. Their story covers the themes of general interest to satisfy the curiosity of the visitors. However, there is not enough time to explain the topics in depth.

As an additional service, the visitors are encouraged to send a text message during the tour from their mobile phone to a specified telephone number with keywords expressing their interests. The text message may contain up to nine keywords that are, together with the phone number, displayed on the labels of the exhibits. The visitor's text message is used as a set of seed concepts. Later at home, the sender can login to the Bletchley Park Text page using the number of his/her mobile phone as a user name and explore a personalised set of stories. The overall architecture is shown in Figure 4.
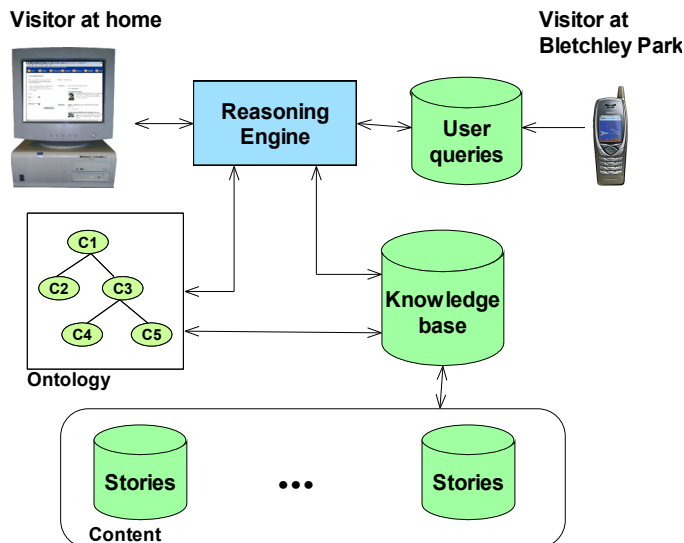
Fig. 4. Architecture of Bletchley Park Text

There are three possibilities of defining pathways in the focused document space:
- only concepts from the focused space are allowed,
- concepts from the focused space are preferred, and
- any concept is allowed.

Option (a) brings the attention of the user to the boundary of the focus while options (b) and (c) introduce the user to new concepts and new themes.

V. EXPLORATION BASED ON INFORMATION CRITERIA

Each exploratory step along the path in the document space is associated with acquiring new information. In each step, the information shared by two events can be measured by mutual information:

$$I(c_i : c_j) = \log_2 (P(c_i, c_j))/(P(c_i) \cdot P(c_j))$$

where concepts c_i and c_j occur with probabilities $P(c_i)$ and $P(c_j)$, and $P(c_i, c_j)$ is the joint probability of concepts c_i and c_j.

The "tighter" relation between concepts c_i and c_j, the higher value of joint probability and the higher value of mutual information $I(c_i : c_j)$. If concepts c_i and c_j are independent, the joint probability equals the product of the individual probabilities and the mutual information is zero. In the document space probabilities can be estimated as relative frequencies. Hence, probability $P(c_i)$ is calculated as $P(c_i) = |E(c_i)| / |E|$, where $|E(c)|$ denotes the number of events that contain concept c and $|E|$ denotes the total number of events of the document space. Similarly, we calculate the joint probability as $P(c_i, c_j) = |E(c_i \& c_j)| / |E|$, where

$|E(c_i \& c_j)|$ is the number of events that contain both c_i and c_j. Mutual information $I(c_i : c_j)$ computes information shared by adjacent documents along the pathway. This information criterion can quantitatively evaluate paths and can be applied both to the exploration of the complete document space and to the focused document space.

A. Example

A simple example of pathways between two concepts in the complete Bletchley Park document space is shown in Figure 5. There are six shortest paths of length four from concept BTM-Works to Colossus-Mk-II. They are shown as a graph (not to be confused with the hypergraph defined earlier) with edges between concepts representing the binary relation *being-member-of-the-same-event*.
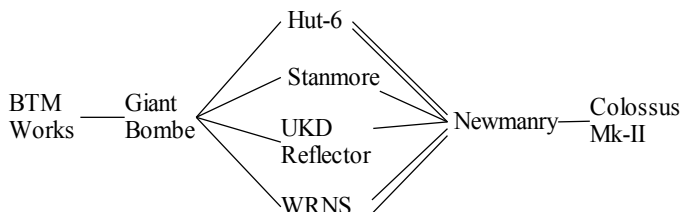

Fig. 5. Paths between BTM-Works and Colossus-Mk-II in the complete document space

Two lines between concepts mean that there are two different events/stories connecting these concepts.

Mutual information of for each step scores as follows:

BTM-Works → Giant-Bombe	$I = 8.589$ [bit]
Giant-Bombe → Hut-6	$I = 3.065$ [bit]
Hut-6 - Newmanry	$I = 0.895$ [bit]
Newmanry - Colossus -MK-II	$I = 5.419$ [bit]

The user may start their exploration of content by reading these stories to better understand what the concepts mean. He/she will find out that: BTM-Works stands for the British Tabulating Machine Company which later became the computer manufacturer ICL Limited; Giant-Bombe was one of the decoding machines; Hut-6 was the workplace in Bletchley Park, Newmanry was a section of codebreakers lead by Max Newman; and Colossus-MK-II was a later version of the Colossus machine.

Let us see what we can infer from the values of mutual information. When we compare the values along the pathway we observe a noticeable dip in the Hut-6 - Newmanry step. This peculiarity deserves further exploration of the domain. It means that stories linking Hut-6 and Newmanry are not very numerous and therefore the joint probability $P(Hut-6, Newmanry)$ has a small value. But, why do only a few stories write about both Hut 6 and the Newmanry? Let us summarise what we could find out about the domain by reading the available documents: Hut 6 was the workplace of staff working on the interception and breaking of Enigma messages using Bombe machines, while the Newmanry team

used a different technology (Colossus) to break more complex Lorenz code. The likely explanation is that the stories are mainly transcripts of interviews with Bletchley Park support staff who had only limited knowledge about activities carried out outside of their workplaces. For security reasons, social contacts between different groups were not recommended and those who worked in Hut 6 had little knowledge about the Colossus technology used in the Newmanry and vice versa.

B. Information based criterion in a focused space

Focusing the document space affects the values of mutual information along the path. The number of events is not reduced evenly across the document space. In particular, the number of events in sets $E(c_i)$, $E(c_j)$ and $E(c_i \& c_j)$ does not change, but focusing removes unrelated concepts and therefore reduces the total number of events from E to E_F, see figure 6.
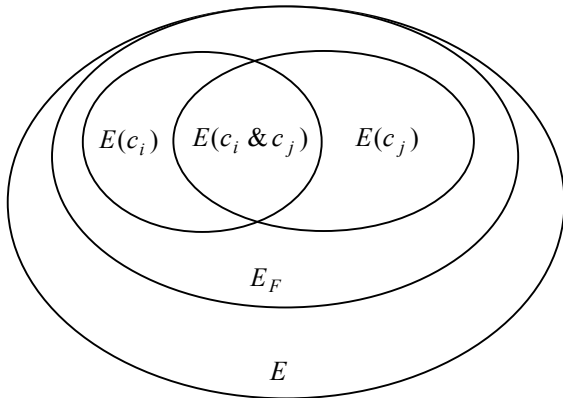

Fig. 6. Reducing document space by focusing

In the focused document space probabilities are $P_F(c_i) = |E(c_i)| / |E_F|$, $P_F(c_j) = |E(c_j)| / |E_F|$ and $P_F(c_i, c_j) = |E(c_i \& c_j)| / |E_F|$. Consequently, mutual information in the focused document space is $I_F(c_i : c_j) = \log_2 P_F(c_i, c_j) / (P_F(c_i) \cdot P_F(c_j))$. By comparing $I(c_i : c_j)$ and $I_F(c_i : c_j)$ we get $I(c_i : c_j) = I_F(c_i : c_j) + \log_2 |E| / |E_F|$. The difference $\log_2 |E| / |E_F|$ is the information gained by focusing the document space, e.g. by stating that stories from $E - E_F$ will not participate in any pathway.

C. Example continued

We will demonstrate the effect of focusing by further extending the previous example. Let's assume that the selected set of seed concepts is S = {BTM-Works, Alan-Turing, COLOSSUS-MK-II}. By applying this seed, the focused document space is reduced from 770 to 314 events. Concepts *Stanmore*, *UKD-Reflector* and *Wrns* do not exist in the new focused space and therefore the paths through these concepts cannot be constructed. The number of shortest pathways has decreased from 6 to 2. They are shown in figure 7 by thick lines.
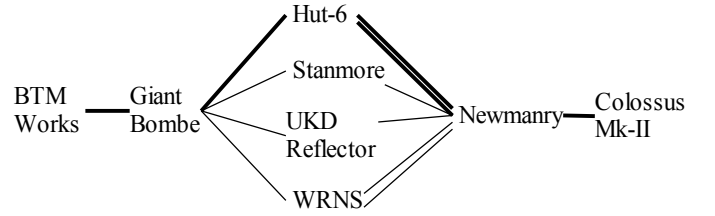

Fig. 7. Paths between BTM-Works and Colossus-Mk-II in the focused document space

Steps in the focused document space have the following values of mutual information:

BTM-WORKS → GIANT-BOMBE	$I = 7.295$ [bit]
GIANT-BOMBE → HUT-6	$I = 1.771$ [bit]
HUT-6 → NEWMANRY	$I = 0.399$ [bit]
NEWMANRY → COLOSSUS-MK-II	$I = 4.125$ [bit]

Focusing eliminates some of the events that were taken into account in the original document space. This reduction can be measured by its information value. The information gained by focusing is calculated as the difference between the corresponding pairs of steps, i.e. 8.589 - 7.295 = 3.065 - 1.771 = … = 1.294 [bit]. The value is in accordance with our calculation introduced in section 5.2.

VI. RELATED WORK

Using pathways for navigation through a set of documents has a long tradition in the hypertext community though its origins can be traced to the memex of Vannevar Bush. In the eighties and nineties the effort was mainly focused on using pathways for easier navigation through hypertext [16]. Interconnected documents formed too complex information networks and users often suffered from cognitive overload and *disorientation*. Even though a hypertext was created by a single author it was very difficult to get the intended message across to the reader. To facilitate browsing, hypertext structures were complemented by some form of recommended sequence of steps. These pathways had different names: guided tours [5], [7], [15]; trails [9] or Walden's paths [12]. However, their purpose was very similar: they order documents into the sequence preferred by the author. In the educational context, the pathways were an instrument of the teacher for organizing the content so that the students do not miss any important facts. The teacher could reuse content of different authors and complete by her own comments and explanations as annotations associated with individual steps in the pathway [12]. In these examples pathways are used to help the reader manage the document collection, but the process can hardly be called exploration.

The World Wide Web (WWW) aggravated the navigation problem by creating hypertext structures from a large number of documents. New models of navigation for the WWW have been developed in [7] and [9]. Both models recognize that the value of a pathway is greater than the values of individual documents and both create the best pathway by evaluating existing static hyperlinks. WebWatcher [7] follows the user

when browsing, proposes the next step and learns from the behaviour of previous users. The model described in [9] represents the hypertext as a probabilistic finite state machine and trail (pathway) as a word of the language accepted by this machine. A user query is a set of keywords and the trail is a sequence of documents such that each document contains at lease one query keyword. For a given user query, the *Best Trail* algorithm evaluates the most relevant trails and provides advice to the user. In this sense, the Best Trail is similar to the pathways in the focused document space. However, the Best Trail makes use of predefined static links while our method constructs links dynamically.

Measuring similarity between documents in terms of mutual information has been used by many authors, e.g. [3]. However, we do not calculate mutual information from the whole document, but only from its annotation and unlike other authors we use this criterion for evaluating steps in the pathway.

VII. Discussion

In this paper we present methods for creating pathways in the document space and using the pathways as a means to support the exploration of the document set. The pathways are not calculated directly from the documents. In the first step we have annotated documents in terms of underlying ontologies. The annotation simplifies the calculation because only relevant concepts and relations are taken into account. The presentation of the resulting pathway can use the annotation for summarizing stories. On the other hand, the process of annotation does not easily scale up. Though we have developed a number of knowledge editors and support tools for checking consistency, completeness and correctness of the resulting knowledge base, the main burden is still carried out by a human annotator. We expect that the acceptable compromise will lie somewhere between the automated processing of raw documents and the annotation currently used by our algorithms. Examples of such lightweight annotation approaches have been described in [13] and [14] and we have also carried out successful experiments.

In the example presented in this paper we use documents which describe structured stories. Though this is not a serious constraint because documents frequently are stories, we have also annotated documents which did not have the story structure. In the Bletchley Park Text application, there are facts that cannot be found in any interview because they were well known and therefore the interviewee would not have cause to mention them. An example is the fact that Newmanry was a group led by Max Newman. These facts are also formally represented in our system and used for reasoning.

Focusing allows the user to express his/her interest and by providing additional information to the system restrict the search space.

Information criterion introduces an additional description of the domain. A low value of mutual information means that the adjacent documents are only loosely related. The reader may expect new insights into the topic, surprises or at least many new concepts. Such a pathway corresponds to an

"adventurous" strategy of content exploration. A high value indicates that the topic is further elucidated with a large number of repeating concepts. Such a pathway corresponds to "cautious" exploration. Information criterion associated with the pathway can serve for guiding the content exploration, i.e. the user can select adventurous or cautious strategy, for informing the user about interesting steps in the pathway or for post-mortem evaluation user's behaviour.

The complete Bletchley Park Text system, including the exploration methods described in [10], is routinely used by Bletchley Park visitors since April 2005.

References

[1] Berge C.: Graphs and Hypergraphs. North Holland. Amsterdam, London., and American Elsevier, New York. (1973).

[2] CIDOC: CIDOC Conceptual Reference Model (CRM), proposal for "ISO 21127: A Reference Ontology for the Interchange of Cultural Heritage Information", http://cidoc.ics.forth.gr/index.html .(2004)

[3] Church K.W. and Hanks P.: Word Association Norms, Mutual Information, and Lexicography. Computational Linguistics, Vol.16. March 1990. pp. 22-29.

[4] Fox E.A., Das Neves F., Yu Xiaoyan, Shen R. Kim S. and Fan W.: Exploring the Computting Literature with Visualisation and Stepping Stones & Pathways. Communications of the ACM. April 2006/Vol.49. No.4. pp. 52-58.

[5] Guinan C. and Smeaton A.F.: Information retrieval from hypertext using dynamically planned guided tours. In Proceedings of European Conference on Hypertext (ECHT'92), (D. Lucarella et al., eds.), (1992) pp 122-130, Milan, Italy.

[6] Hearst M.A.: Clustering versus Faceted Categories for Information Exploration. Communications of the ACM. April 2006/Vol.49. No.4. pp. 59-61.

[7] Joachims T., Freitag D. and Mitchell T.: WebWatcher: A Tour Guide for the World Wide Web. In Proceedings of the 15th IJCAI '97, Nagoya, Japan. (1997), pp. 770-775

[8] Landow G.P.: Hypertext 2.0. John Hopkins University Press. Baltimore. (1997).

[9] Levene M. and Wheeldon R.: Navigating the World-Wide-Web. In Web Dynamics,Adapting to Change in Content, Size, Topology and Use. (Levene M. & Poulovassilis A. eds.), (2004). pp. 117-151

[10] Mulholland P, Collins T. and Zdrahal Z.: Spotlight browsing of resource archives. In HT 2005 (Sixteenth ACM Conference on Hypertext and Hypermedia)**,** 6.-9.Sept. (2005). Salzburg, Austria

[11] Schank R.C.: Tell me a story. Northwestern University Press, Illinois. (1990).Third printing 2000.

[12] Shipman F.M. III, Furuta R., Brenner D., Chung C. C. and Hsieh H. W.: Using Paths in the Classroom: Experiences and Adaptations. In ACM Hypertext '98 Proceedings, (1998). pp. 267-276

[13] Shipman F M.III, and McCall, R.: Incremental Formalization with the Hyper-Object Substrate. ACM Transactions on Information Systems (1999). Vol.17, Issue 2. pp. 199-227

[14] Shipman, F.M. III, Moore, J. M., Maloor, P., Hsieh, H., and Akkapeddi, R.: Semantics happen: knowledge building in spatial hypertext. In Proceedings of the Thirteenth ACM Conference on Hypertext and Hypermedia (College Park, Maryland, USA, June 11 - 15, 2002). HYPERTEXT '02. ACM Press, New York, NY, pp. 25-34.

[15] Trigg R.H.: Guided tours and tabletop: tools for communicating in a hypertext environment, ACM Transactions on Office Information Systems. (1988).Vol. 6. No.4. pp. 398–414.

[16] Zellweger P.T.: Scripted Documents: A Hypermedia Path Mechanism. In Proceeding Hypertext'89. (1989). pp. 1-14.