# Exploring Pathways Across Stories

Zdenek Zdrahal, Paul Mulholland, Trevor Collins
Knowledge Media Institute, The Open University,
Walton Hall, Milton Keynes, MK7 6AA, 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

## Introduction

This paper describes a set of methods for the exploration of a collection of stories. The term "story" refers to a semantically self-contained reading unit (i.e. lexia) comprising a block of text with associated pictures or multimedia. There are two main reasons for concentrating on documents in the form of stories: knowledge is often represented in stories and there is a natural way of breaking the document into smaller units (i.e. events). We assume that the documents in the collection are semantically interrelated, that is, they share common key concepts and refer to related events. Therefore, the whole collection can be regarded as a form of hypertext. However, unlike a standard hypertext, the links between lexias are not explicitly predefined, but are dynamically constructed in accordance with the user's interests from the documents' annotations. In addition to the knowledge extracted from individual documents, further meaning can be inferred from organizing and presenting documents in different structures, and in this way facilitate the discovery of the knowledge hidden in the collection.

## Case study: Bletchley Park

During World War Two, Bletchley Park was 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 and the enlisted tour guides started collating information about the history of Bletchley Park. At present, the archive consists of thousands of unique documents about: code breaking, early computing, life and work of prominent scientists and other staff in Bletchley, the impact of the Bletchley Park effort on the course of the war and other similar topics.

### Bletchley Park Text

The application we have developed for the 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 used as the upper ontology (CRM, 2004). The next level ontologies include the *bletchley-park-ontology* which specializes the CRM 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 dynamic links between the annotated documents (stories or events) are defined in terms of *binding concepts* which are in this case instances of classes *actor*, *place* and *object* stored in knowledge bases. An example story annotation is shown in Figure 1.

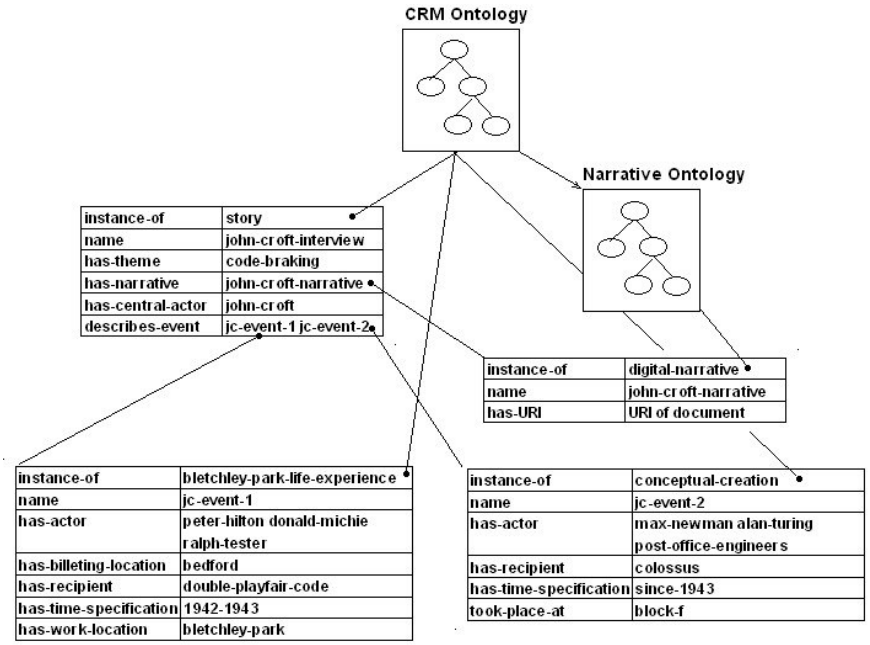


Figure 1 Story annotation showing slots and their values

A pathway is a sequence of stories in which two adjacent stories share a binding concept. Slot types are ignored for defining paths, but they are used later for interpreting paths.

## Formal representation

The document space can be represented as a hypergraph $H = <C,E>$, where $C = \{c_1, \ldots, c_N\}$ is a set of nodes corresponding to concepts and $E = \{e_1, \ldots, e_M\}$ is a set of edges corresponding to events (stories, documents).

The document hypergraph is constructed as follows:

1. Annotated events in all documents specify the set of edges $E = \{e_i\}$. Edges are n-tuples of concepts with associated event names for easy identification.

2. The set of nodes $C = \{c_j\}$ is defined as a union of all edges, $C = \cup e_i$. The slot names of events are not used.

Fig. 2 shows a hypergraph with 13 annotated concepts and 5 documents.

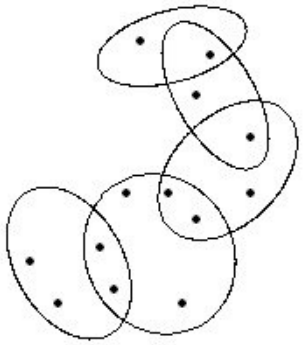

Figure 2 Representing document space by a hypergraph

## Exploration of pathways

What can the user achieve by constructing and investigating pathways in the document space? Arranging documents into a meaningful sequence makes it possible to discover new information. For example, it is possible to deduce that person A may have known person B (e.g. if they were in the same place at the same time), person A may have heard about technology C (e.g. if there is a chain of events from person B who knew about technology C to person A and this chain allowed information about C to be transferred). The analysis of mutual information along the pathway (which is described later), will allow us to draw interesting conjectures about life in Bletchley Park.

In this case we consider only shortest pathways in the document space, defined in terms of the number of connected concepts. Even with this constraint, there are usually multiple pathways between any two concepts. Depending on their properties, they contribute in a different way to content exploration.

### Simple pathway between two concepts

Conceptually the simplest is a pathway connecting two concepts in the document space without any 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 $k$.

### 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 characterize the themes of interest. The seed concepts could be any binding concepts (e.g. people, places, objects) 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 $S = \{c_1, \ldots, c_s\}$ 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 $H$ and the corresponding subspace of the original document space is called *focused document space*. Focused document space is shown in Figure 3.

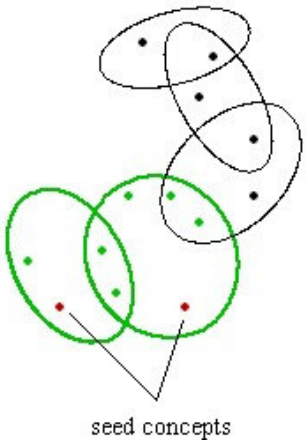


Figure 3 An example focused document space with two seeds identifying eight concepts from two stories.

Focusing reduces the document space to be explored. If all seed concepts belong to the same theme, then the focused document space contains only concepts of this theme and paths can be constructed only from concepts, events and stories of this theme. However, as the themes are overlapping seed concepts may select multiple themes. If seed concepts hit themes represented by disjoined clusters the focused space becomes disconnected and there are concepts in *CS* for which a pathway in *HS* does not exists.

## 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 any specific 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 Fig. 4.

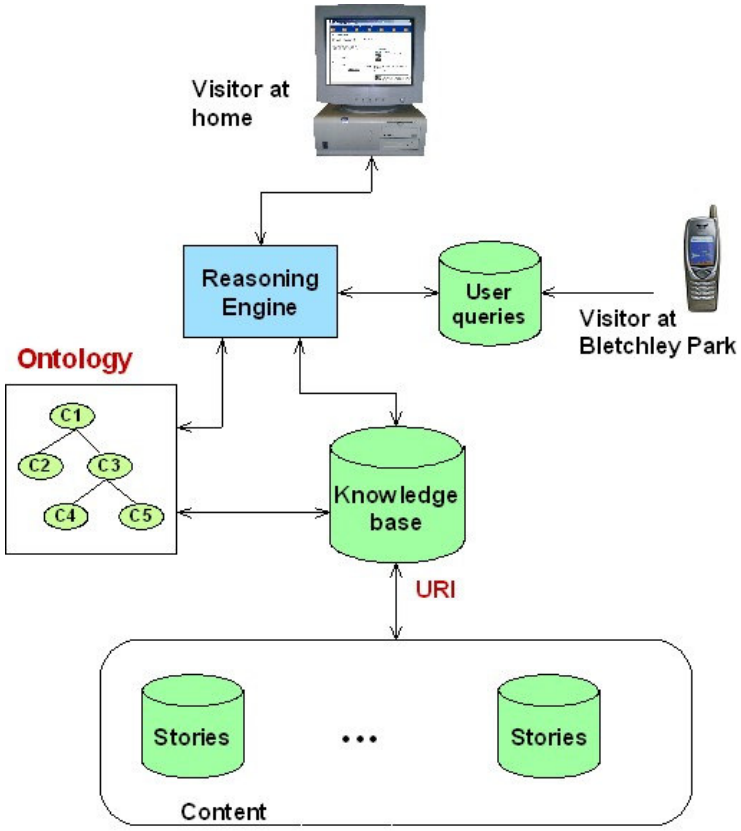


Fig. 4. Architecture of Bletchley Park Text

There are three possibilities of defining pathways in the focused document space:

- (a) only concepts from the focused space are allowed,
- (b) concepts from the focused space are preferred, and
- (c) any concept is allowed.

Option (a) brings the attention of the user to the boundary of the focus while option (b) introduces the user to new concepts and new themes.

## Mutual information along pathways

Each step along the path is associated with acquiring new information. In each step, the information shared by two events can be measured by mutual information $I(c_i : c_j)$, defined as

$I(c_i : c_j) = \log_2 (P(c_i, c_j)/(P(c_i) \cdot P(c_j)))$, where $P(c_i)$, $P(c_j)$ and $P(c_i, c_j)$ are the probabilities of $c_i$, $c_j$ and a joint probability of $c_i$ & $c_j$, respectively.
The "tighter" the relation between concepts $c_i$ and $c_j$, the higher the value of the joint probability and the higher the value of mutual information $I(c_i : c_j)$. If concepts $c_i$ and $c_j$ are independent, the joint probability is equal to the product of the individual probabilities and the mutual information is zero. In the document space, probabilities are estimated as relative frequencies, i.e. $P(c_i)$ is calculated as $P(c_i) = |E(c_i)|/|E|$, where $|E(c)|$ denotes the number of events associated with concept $c$ and $|E|$ denotes the total number of events in the document space.

## Example

An example of the shortest pathways between two concepts is shown in Figure 5.

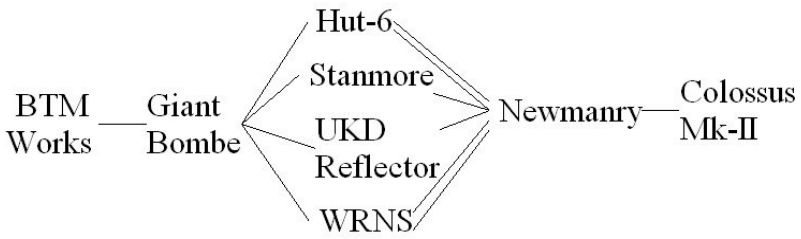


Figure 5. Shortest paths between BTM-Works and Colossus-MK-II

There are 6 shortest pathways of length 4 between concepts BTM-Works and Colossus-Mk-II. There are two alternatives for the step from Hut-6 to Newmanry and from WRNS to Newmanry.

An example of the pathway through Hut-6 is:

BTM-Works - <JAN-44-IU-2> - Giant-Bombe - <JAN-44-IU-3> - Hut-6 - <JAN-44-P5L-1> - Newmanry - <APR-44-OFC> - Colossus-MK-II,

where <JAN-44-IU-2>, <JAN-44-IU-3>, <JAN-44-P5L-1> and <APR-44-OFC> are the names of events. The mutual information scores of each step are 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 explore the content by reading these stories to better understand what the concepts mean. Let us see what we can infer from the above 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 analysis of the domain.
It means that there are only a few stories about Hut 6 and the Newmanry. But why? 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 workplace. Hut 6 was the workplace of staff working on breaking Enigma codes using the Bombe machine, while the Newmanry team used the Colossus computer to break the more sophisticated Lorenz codes. For security reasons, social contact between different groups were discouraged and therefore those who worked in Hut 6 had little knowledge about the Newmanry and vice versa.

## 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 at all, while focusing removes unrelated concepts and therefore reduces the total number of events from $E$ to $E_F$, see Figure 6.

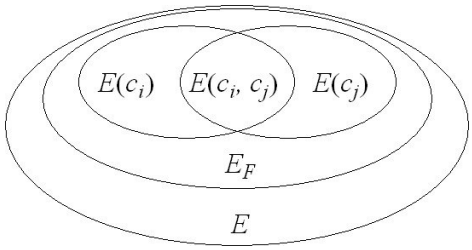


Fig. 6 Reducing document space by focusing

In the focused document space probabilities change the value to $PF(c_i) = |E(c_i)|/|EF|$, similar formulae hold for $PF(c_j)$ and $PF(c_i,c_j)$. It is easy to prove that by focusing the mutual information of each step is reduced by $\Delta I = \log_2 |E|/|E_F|$. The difference $\Delta I$ is the information gained by focusing, i.e. by stating that events $E - EF$ can be ignored.

**Example (cont.)**

By applying the seed S = {BTM-Works, Alan-Turing, COLOSSUS-MK-II}, the document space is reduced from 770 to 314 events. Concepts Stanmore, UKD-Reflector and WRNS do not belong to the new focused space which eliminates 4 pathways. The remaining 2 pathways are highlighted in Figure 7.

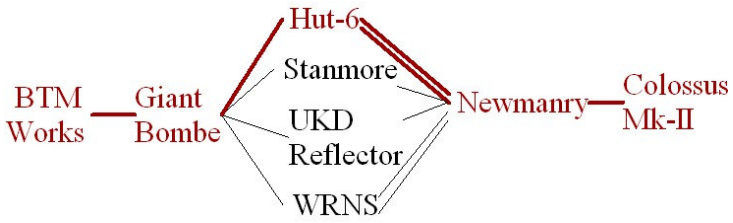


Fig. 7 Pathways in the focused document space

In accordance with the formula for $\Delta I$, focusing decreased the mutual information of each step by 1.294 [bit].

## Conclusions

The demand for efficient knowledge management techniques is growing due to the increasing complexity of large information networks (Fox et al. 2006). Pathways offer one possible solution for avoiding the cognitive overload and disorientation of the user (Levene & Wheeldon, 2004). Organising documents into a sequence connected by common concepts allows the user to recognize and infer the information scattered across many documents. In this paper we have presented a number of methods for constructing and ranking pathways. Having the learning scenario in mind, the described criteria take into account whether the concepts along the pathway are familiar or new to the user, whether they unfold new themes in the domain, and whether the information is provided in small steps or big leaps. The described techniques have been implemented in the Bletchley Park Text system which has been used by visitors to Bletchley Park since May 2005. For further information including a demo see http://kmi.open.ac.uk/projects/bp-text.

## References

CRM (2004). 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

Fox E.A., Das Neves F., Yu Xiaoyan, Shen R. Kim S. and Fan W. (2006). Exploring the Computting Literature with Visualisation and Stepping Stones & Pathways. CACM. Vol.49. No.4. pp. 52-58.

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