Dynamically Detecting and Modelling the Visitor’s Interests in a Museum Environment

A thesis presented
by
Michael Yang
to
The Department of Computer Science and Software Engineering
in partial fulfillment of the requirements
for the degree of
Bachelor of Computer Science (Hons)

University of Melbourne
Melbourne, Australia
October 2007
Dynamically Detecting and Modelling the Visitor’s Interests in a Museum Environment

Abstract
A museum contains a diversity of information sources, from which a visitor can select according to his interests. Although every visitor may have a different purpose for visiting the museum, in our research we assume the motivation for all visits is to get information. However, more often than not, visitors do not obtain a thorough coverage of information on the topics of their interests, given a limited time duration for the tour. To maximise this coverage, the solution is to increase our knowledge of each individual’s information needs, and develop a system to provide personalised and contextually relevant navigation to the visitor. The personalisation component ensures the provided information is adaptive to the interests of each unique visitor, whereas the contextual relevance component looks at the state of the visitor in relation to their surroundings. In this project, we develop a prototype system that aims to address both components; in particular, we experiment with various computational linguistic models on the prototype system and contrast their effectiveness in detecting and modelling the visitor’s interests during the museum tour.
Contents

Title Page ................................................................. i
Abstract ................................................................. ii

1 Introduction ......................................................... 2
   1.1 Task Background ................................................ 3
   1.2 Methodology ..................................................... 3
   1.3 Overview ........................................................ 4

2 Background and Previous Work .................................. 5
   2.1 Integrating Physical and Virtual Environment .............. 5
   2.2 Navigation in the Museum ..................................... 6
   2.3 User Modeling and Interaction ................................ 7

3 Resources .......................................................... 8
   3.1 Museum Data .................................................... 8
      3.1.1 Melbourne Museum ....................................... 8
   3.2 Natural Language Resources ................................ 11
      3.2.1 Natural Language Toolkit ............................... 11
      3.2.2 WordNet .................................................. 12
      3.2.3 SemCor .................................................... 14
      3.2.4 SENSELEARNER ......................................... 14

4 Methodology ........................................................ 16
   4.1 Content Manipulation ......................................... 16
      4.1.1 Term Weighting .......................................... 17
      4.1.2 Semantic Depth .......................................... 18
      4.1.3 Word Sense Disambiguation ............................ 20
   4.2 Computational Models ......................................... 20
      4.2.1 Word-Level Model ....................................... 21
      4.2.2 Uniform tf.idf Weight Distribution .................... 23
      4.2.3 Most Frequent Sense .................................... 24
      4.2.4 Learnt-Sense ............................................. 25
      4.2.5 Summary .................................................. 26
Chapter 1

Introduction

People often claim that “there is too much data, but little information”. This is true in the sense that, despite the large quantity of data in a given environment, people often find it hard to gather data that is truly useful and relevant to them. If people’s needs and interests can be known in advance, then information providers could cater to the possible requirements of each individual, and hence deliver more personalised services to a user\(^1\) by purposely selecting information from large repositories of data. However, there is often very limited prior knowledge of users’ needs in most environments. Even if such needs can be known in advance, there are always possibilities for them to change unpredictably. Besides, if we were to use simple word tokens to represent a visitor’s interests, say, then words such as *cricket* and *plane* may refer to completely different objects or concepts depending on their context.\(^2\) We would then have no information as to whether a given visitor is interested in cricket as a sport or an insect. Therefore, static, syntax-level knowledge might not be fully reliable in modelling the user. This forms the motivation for our research to ensure that the detection and modelling of a visitor’s interests are both dynamic and context-sensitive.

For the purposes of our research, the test environment is chosen to be a museum. For any museum, it is possible that different visitors’ interests cannot be fully categorised in advance. However, they can be inferred from visitors’ behaviour and context,\(^3\) e.g. the visitor’s locations, path history, etc. In this research, we aim to build a prototype system that can help us make such inferences about the visitor’s interests. We implement various computational models, and examine how each of them impacts on the modelling of the visitor’s interests, in relation to how information is represented within the museum space.

\(^1\)The term ‘user’ refers to a museum visitor in the context of our research

\(^2\)The context of a word is the overall background or topic that covers this word.

\(^3\)The context of a visitor refers to his current state in the museum. (The word *context* itself depends on the context!)
1.1 Task Background

Our research forms a part of the Kubadji Project,\textsuperscript{4} which is a collaborative project with Melbourne Museum. The Kubadji Project investigates computational models to provide personalised information to assist a visitor in learning and interacting within the museum. It involves the automatic generation of personalised information to the visitor, either relating to a visited exhibit or recommending an unseen exhibit of the visitor’s interests. Our research does not focus on the details of information delivery to the visitor, nevertheless, it contributes to the learning of a visitor’s interests, which lays the basis for generating personalised information. In the broad vision, our research pertains to the fields of user-modelling and context-aware computing, where new approaches are needed for acquiring, representing, updating and exploiting models of an active user in a dynamic, information-rich environment (Jameson and Kruger 2002). To date, combining contextual information with visitor models still remains a key challenge for personalised information systems (Zimmermann \textit{et al.} 2005).

The main areas of focus for our research are: User Modelling – simulating and inferring the interests of each individual visitor relative to the museum environment; and the application of Computational Linguistics – using Natural Language Processing (NLP) techniques to compute various representations of information in the museum environment. It should be noted that our research does not focus on the prediction of the visitor’s path, and we do not pay attention to aspects related to the physical locations of exhibits and visitors. However, our research is compatible with both cases, and they are all contributive to user-personalisation overall.

1.2 Methodology

The reason we choose a museum as our test environment is mainly due to its diversity of information sources, which can help us identify the interests of a visitor based on what information he has been exposed to. Due to the correspondence between the museum’s online documents and the physical exhibits, we use the textual document as our source of information for each exhibit. From each document, our system extracts content words and use certain linguistic features of these words to assist in the modelling of the exhibit content. As the visitor stops at a given exhibit, the modelled content of this exhibit can be used to infer the visitor’s interests at that point in time. This allows us to update our knowledge of this particular visitor, hence being adaptive to his information needs at any time, any place during the tour. This creates a more dynamic notion of user-learning than simply relying on his preferences prior to the tour.

\textsuperscript{4}http://www.kubadji.org/
Chapter 1: Introduction

With the content words harvested from the museum documents, we explore their features such as word senses, synonyms, and degrees of specialisation in meaning. Using these linguistic features, along with tools including WordNet (Yarowsky 1992) and SenseLearner (Mihalcea and Faruque 2004), we design and implement four computational models, each of which has a unique method in representing the content of exhibits. As the content of exhibits is modelled differently, the inference of a visitor’s interests would then vary between each model, depending on what exhibits he has seen, and how we derive potential areas of interest out of them. Hence, under each computational model, the prototype system uniquely learns about the visitor, and infers a list of the visitor’s interests at the end. This list contains information about what each model thinks the visitor is interested in throughout the tour, and will be compared against our gold standard, i.e. visitor’s feedback on what each visitor is indeed interested in. We will contrast the performances of all four models, and analyse how they are influenced under various methods of linguistic computations.

1.3 Overview

Different aspects of user modelling and language technology form the basis of our research. We will begin by first outlining the relationships between these fields of research and describe a number of research work conducted previously in these fields. (Chapter 2). This is followed by a discussion of resources used in this research, including the text data from the museum webspace, as well as a number of language technology tools used to manipulate this text data (Chapter 3). The exploration of multiple computational models and their effect on user-modelling is the core contribution of this research (Chapter 4). An in-depth analysis of our results contrasts the effectiveness of each model in its inference of the visitor’s interests. (Chapter 5). We end with an overall reflection of our research, and provide discussions for the future work (Chapter 6).
Chapter 2

Background and Previous Work

Our research is part of the Kubadji Project that attempts to develop a system to deliver personalised information to visitors for navigation purposes during their tour of the museum. The project explores an integration between physical exhibits and their corresponding documents, attempting to extract information from the document space and present it to a given visitor based on the exhibits seen. Hence, the information delivered is targeted at each specific visitor, and is designed to draw conceptual links between exhibits. This chapter describes a number of previous projects that tend to investigate various fields of research that are similar to our work.

2.1 Integrating Physical and Virtual Environment

There have been a number of major projects around the world focusing on the integration of features from the physical and virtual (online) world. HIPS – Hyper Interaction within Physical Space – is one such project (Benelli et al. 1999). The concept behind HIPS extends the notion of hypernavigation, that is, it not only targets navigation within a virtual environment but also aims to develop interaction modalities that allow people to simultaneously navigate both a physical space and a related information space with a minimum gap between the two. The overall goal of the HIPS Project is very similar to the Kubadji Project: both involve the creation of a portable electronic tour guide for the purpose of recommending physical objects to each visitor according to his own criteria. The difference is that we investigate various ways of manipulating information content from a linguistic angle, exploiting the application of language technology to incrementally learn about the visitor as he tours through the museum. HIPS, on the other hand, focuses more on the physical location of each object in order to infer its information relative to the surrounding objects, which can then be recommended to the user.
The Equator Project\textsuperscript{1} is another major project in the same broad research field. It is an interdisciplinary research collaboration that explores and addresses the challenges of the cooperation between digital and physical activities. Its core contribution is to incorporate the analysis of how objects in the digital environment (such as the Web and GPS technology) can create meaningful correspondence with objects in the physical environment. Randell \textit{et al.} (2006) demonstrated that it is possible to analyse human factors together with location-based aspects, and combine them to give a more comprehensive understanding of the factors that influence performance. One such human factor examined in our research is the time duration spent at each exhibit in the visitor’s path (Section 4.2.1).

Not and Zancanaro (1998) define physical hypernavigation as a user-personalised tour, making use of the synchronisation between the physical space and informational hyperspace. Similar to the Kubadji Project, this synchronisation is carried out by selecting and organising meaningful information to be displayed on a mobile device carried by each visitor during the tour. This is also the vision of the PEACH Project (Kuflik \textit{et al.} 2004), which aims to provide educational entertainment to visitors through their personal experiences in the museum. PEACH shares many similar features with our research, in that they both intend to link the physical space, where the objects are located, with the information space, where the meaning or the descriptive content of the objects is located and pre-processed, and hence language technology plays an important role in both projects.

The software prototype we develop in our research can be integrated with hardware components such as Radio-Frequency Identification (RFID) tags, which detect and track the positions of the visitor, thus providing data about the visitor’s context. For the purposes of our research, we do not focus on the hardware side of the project, neither do we pay attention to the prediction of the visitor’s path, as we stated explicitly in our introduction. Instead, we concentrate on how to dynamically infer individuals’ interests from their context, in associations with various ways to represent information content of exhibits.

2.2 Navigation in the Museum

As the visitor tours through the museum, his path can be envisaged as a collection of information about the exhibits visited. Given the approximate one-to-one correspondence between each exhibit and an online document, we represent each document as a ‘collection of semantic words’ or a ‘bag of words’ (Freitag 1998), which is in turn indicative of the information conveyed from the exhibit. For example, a document corresponding to the exhibit of a grass frog might contain words such as \textit{cricket}, \textit{river},

\textsuperscript{1}http://www.equator.ac.uk
and *amphibian*. Intuitively, these words describe the concepts of food, habitat, and species of the frog, respectively. The challenge is, for an ambiguous word such as *cricket*, we need to artificially identify that it refers to an insect rather than a type of sport, given that our context is about animals. Having said this, if another document describing a *toad* also happens to contain the word *amphibian* in its content, then according to Grieser (2006), we can treat the frog and the toad as two separate exhibits that share a similar concept, namely, *amphibian*. Hence, if the visitor looks at both exhibits, we can infer that this visitor could be interested in amphibians. This inference then enables us to recommend another amphibian-related exhibit to this visitor, thus providing personalised navigation in the museum.

Intelligent Labelling Explorer (ILEX) (Hitzeman et al. 1997) is another large research project that shares some common concepts with our research. In the ILEX Project, given that every exhibit in the museum corresponds to a webpage (though it is not strictly the case in our research), as the user clicks through webpages, the system is able to generate descriptions of corresponding exhibits, hence giving a sense of a simulated tour in the museum. This generated description of each exhibit is based on identifiable aspects of the exhibit (Coxa et al. 1999) such as keywords and semantic phrases. For example, as a visitor sees ants, a spider, and then a turtle, the ILEX system keeps a log of the visitor’s history, based on which the system then tries to relate to unseen exhibits, while generating text to explain the conceptual relationships between ants, the spider, etc. to help the visitor understand more about what he sees. The nature of the ILEX is very similar to the Kubadji Project in that they both involve the generation of text to enhance the visitor’s experience during the tour (Grieser 2006). In our project, however, we do not investigate the generation of text.

### 2.3 User Modeling and Interaction

The *user model* is an abstract representation of a visitor’s interests and background, based on which we can recommend personalised activities to him (Brusilovsky 1996). Zukerman and Albrecht (2001) develop predictive statistical models to simulate certain aspects of human behaviour, such as actions and preferences. As part of such modelling, the observable, low-level features (such as elapsed time in front of an exhibit), and the informational, unobservable features (such as the personal interests and expertise that are high-level) need be identified. Inferring unobservable information about a visitor from low-level observables is the core for formulating the user model. According to Zukerman and Albrecht (2001), there are two main methods for building user models, *content-based* and *collaborative*, where the visitor’s informational features are inferred from his own observable features and previous visitors’ features, respectively. For our research, we aim to develop a truly user-personalised system, thus we exploit the content-based approach.
Chapter 3

Resources

Melbourne Museum is a resource-rich environment due to the quality and quantity of its information. This is the case for both its physical space and website. It is not difficult to find online information associated with each physical object in the museum, where the quality and consistency of data are both of high standards (Grieser 2006). The museum’s webpages are our data source, where each online document within our dataset will be pre-processed. This is carried out by using language processing techniques to exploit the linguistic properties of words in the context of each exhibit. In this chapter, we describe the sources of museum data, as well as the tools for processing this data in our research.

3.1 Museum Data

3.1.1 Melbourne Museum

Melbourne Museum is a large, well-categorised, and accessible collection of information. Its information can take the form of multimedia such as video (e.g. a movie on bushfires) and audio (e.g. a description of Victorian dinosaurs), images, or descriptive text. The information is organised into levels of granularity: the museum itself is composed of various galleries, where the content of each gallery revolves around a general theme, such as Australian Culture, Forest Secrets, and Prehistoric Life. Within each gallery, a number of exhibitions are organised according to sub-themes, such as Climate, Earth, and Humans, in the case of the Forest Gallery. Each exhibition is in turn made up of exhibition areas, such as the Victorian Bush, which can then be subdivided into a group of single exhibits, such as the silver wattle and currant bush, which are atomic units in the museum space. Considering the scale of our research, we chose the Australian Gallery, Forest Gallery, and Prehistoric Life Gallery as our sources of text data. For each of these three galleries, the locations of exhibits are known in advance. Figure 3.1 shows a map of the Prehistoric Life Gallery with labelled exhibits.
Figure 3.1: Map of the Prehistoric Life Gallery, with each exhibit labelled.
Chapter 3: Resources

The Forest Gallery

The Forest Gallery\(^1\) is a mini-zoo with a wide variety of plant and animal species, including mountain ash, acacia, frogs and insects, to name a few. The gallery is divided into five zones (exhibitions): Water, Earth Processes, Climate, Fire, and Human Intervention. Each zone has a one-to-one correspondence with a sub-domain on the museum website, which makes web-crawling (see Chapter 4) relatively easy. One major reason for choosing the Forest Gallery is due to its range of exhibits on geological facts, which is also a major sub-theme in the Evolution Gallery. This is important for the assimilation between exhibits across different galleries, which is expected to influence our modelling of the visitor. From the documents within the Forest Gallery’s webspace, we observed a range of words including *bat*, *cricket*, and *grain*, which are also found in the contexts of sport and gold-rush in the Australian Gallery. This notion of ambiguity will be dealt with in our computational models (Section 4.2).

The Prehistoric Life Gallery

The Prehistoric Life Gallery\(^2\) makes up the lower level of the Evolution Gallery and explores how life changes over time. This gallery also happens to feature exhibits on geological knowledge about Earth, which we also find in the Forest Gallery. Besides, the content of exhibits in this gallery often include highly specialised words such as *dinosaur* and *trilobite*, which can be helpful to model the interests of a visitor, as we will see in Section 4.1.2.

The Australian Gallery

The Australian Gallery is the third gallery involved in our research. The gallery itself does not have an URL, but some of its sub-galleries do, such as Phar Lap.\(^3\) The Australian Gallery is located on the upper level of the museum and is likely to be visited before or after the user visits the Forest Gallery. The fact that both galleries contain information that is specific to Victorian culture means that a list of exhibits from both galleries can be conceptually assimilated and contribute to the detection of visitor’s interests.

Data on the Web

As stated earlier, we extract the content of all exhibits from documents online. The museum website not only provides information about the exhibits in the museum,

\(^1\)http://melbourne.museum.vic.gov.au/exhibitions/gallery/forest.asp
\(^3\)http://www.museum.vic.gov.au/pharlap/
but also describes the areas of research conducted within the museum, as well as educational programs. The museum webpages are structured in the same hierarchical way as the physical environment: webpages that correspond to similar exhibits are grouped together under the same sub-themes, which are then grouped according to the common theme at a higher level. Such alignment produces the backbone of the synchronisation between the online and real-world environment, allowing us to follow the hyperlinks and map each document against its corresponding exhibit. We refer to these online documents as unambiguous documents since each document corresponds to one exhibit only.

Given an exhibit, if we could not manually find an unambiguous document for it, then we utilise the webpages corresponding to educational programs. Each educational webpage can often be segmented into sub-topics, each of which corresponds to one exhibit in the museum. For any document that does not correspond to a distinct exhibit, we refer to it as an ambiguous document. Details of how to handle such documents are discussed in Section 4.1. Furthermore, there also happen to be cases where the search functionality returns no helpful documents, in which case we seek external sources. For example, there is no online resource that corresponds the physical exhibit on index fossils, therefore we query the Wikipedia\(^4\) database and retrieve text on this topic. For the purpose of our research, having external data sources might not be necessarily helpful, due to the unpredictability of their content. Strube and Ponzetto (2006) discuss methods of determining semantic relatedness of two given concepts using Wikipedia, but it needs to be studied in the future to determine how useful it could be for our research.

### 3.2 Natural Language Resources

The core of our research lies in the application of computational linguistic methods to dynamically model the user. We now review the language technology tools used in our research.

#### 3.2.1 Natural Language Toolkit

The Natural Language Toolkit (NLTK) is a suite of language processing modules (Bird and Loper 2004). Given an online document, the most basic text-processing step using NLTK is tokenisation, which is responsible for demarcating sections of the hypertext into word tokens – the most fundamental units of the ‘bag-of-semantic-words’ structure that we use to represent each online document. However, not every single word within the document is indicative of the exhibit content; for example, words such as *the, of, is* etc. are referred to as stopwords and do not carry specific

\(^4\)http://en.wikipedia.org/
meaning in the museum context. We use the SMART list (Cooke et al. 2004) to remove all stopwords from our bag of words. For the remaining word tokens, NLTK has a readily available tool to lemmatise them, which means we restore their root-lexical form, e.g. lemmatising the word well and better would give their root-form: good. This step is important because it organises a set of syntactically-different words into a commonly recognisable format according to their common semantics. This ensures a greater level of consistency in the comparisons between any two words. Furthermore, abstracting away from the syntactic representation to semantic representation is of significant values to text understanding. For the purposes of our research, this is crucial for modelling an exhibit based on its content, based on which we then model the visitor.

Our software prototype is developed using Python since it is also the implementation language for NLTK. The easy accessibility of lexical databases such as WordNet makes NLTK an essential tool for our research.

3.2.2 WordNet

WordNet is a large electronic lexical database of English, where nouns, verbs, adjectives and adverbs are grouped into sets of synonyms or synsets, each expressing a distinct concept (Miller and Fellbaum 1998). As said above, we need to abstract away from words to their semantics in order to accurately represent the information of an exhibit. This is where WordNet is used – it enables us to represent each word as a semantic concept, and hence each document can in turn be envisaged as a ‘bag of concepts’. In our processing of the document content, we only deal with nouns, due to three main reasons: firstly, nouns are commonly used in linguistic research to represent or summarise information content because they are more informative (Resnik 1995); secondly, by observation, nouns and noun phrases make up a large proportion of the museum text collection; lastly, nouns, verbs and adjectives are in separate lexical databases on WordNet, and hence if we were to choose one group for simplicity and consistency throughout our research, we would choose nouns. Therefore, we exclude verbs, adjectives and adverbs overall. The first reason is important for our purpose of extracting or summarising the semantic content from a document (hence nouns are referred to as content words). As a result, we can then attempt to correlate two documents by correlating the content words from each. WordNet’s taxonomical organisation of nouns provides a significant advantage in this case. Using WordNet, we can estimate the semantic similarity between two nouns, given their positions in the taxonomy. For example, if the visitor is interested in frogs, then regardless of whether or not he has seen a toad, WordNet can be used to detect the semantic similarity between the two terms. The reasons are: firstly, the synset of the first word sense for frog includes the word toad, i.e. these two words are synonyms, which means they are conceptually the same or similar (see Figure 3.2); secondly,
Figure 3.2: The word *frog* and its 3 synsets (sets of synonyms).

Figure 3.3: A rough sketch of the 3 synsets of the concept *frog* in WordNet’s hypernym hierarchy. We take \{noun: entity\} as the most generalised concept for our analysis.

Both *frog* and *toad* can be regarded as subclasses of *amphibian* (see Figure 3.3). The assimilation between concepts is based on the fact that synsets are interlinked by means of conceptual-semantic and lexical relations on WordNet (Miller and Fellbaum 1998). These lexical relationships include synonyms, hypernyms (generalisations) and hyponyms (specialisations), all of which are used to explore the conceptual links between content words in different documents.

For instance, Figure 3.2 shows the synsets of the word *frog*. As observed, the word *frog* has three senses, each corresponds to a unique synset, and each sense/synset pair refers to a unique concept. As Figure 3.3 shows, the second synset refers to the hyponym of the concept \{noun: Frenchman, Frenchwoman, French person\}, whereas the hypernym of the third synset is \{noun: adornment\}. Only the first synset is a hyponym of the concept \{noun: amphibian\}, which is intuitively the concept of interest in the museum context. In this way, WordNet presents information about word senses, allowing us to decide on which sense of a particular word is applicable in a given context. The structured network of conceptual links and senses makes WordNet a powerful tool for analysing and manipulating the content of documents. The version used in this project is WordNet v2.1.
3.2.3 SemCor

SemCor is the largest publicly available sense-tagged corpus (Magnini et al. 2002). It was created at Princeton University by the same team who created WordNet. It is composed of documents extracted from the Brown Corpus that were analysed both syntactically and semantically (Landes et al. 1998). The Part-of-Speech tags (i.e. labelling of words as noun, verb, etc. in the given context of the sentence) were assigned by a tagger, whereas the semantic tagging (e.g. given that cricket is a noun, choose its word sense) was done manually, using the synset information from WordNet. By recording the occurrences of semantic tags (senses) of words within the SemCor corpus, WordNet incorporates the frequencies of all senses for each word. This frequency information is utilised by one of our computational models, which says: for all word tokens, we need to use the most frequent sense of each word according to their frequencies in the SemCor corpus. The motivation and details of this model are described in Section 4.2.3.

3.2.4 SenseLearner

In our ‘bag of words’ conceptualisation of a document’s content, the words are taken to be content words, i.e. nouns that are indicative of the exhibit content (Refer to Section 3.2.2). These content words can either be polysemous (having multiple senses, e.g. cricket) or monosemous (having a single sense, e.g. dinosaur). In our earlier example of the polysemous word frog, we need to be cautious about the selection of word senses. For instance, we should not choose the sense {noun: frog, Gaul} in the context of the Forest Gallery, otherwise, if the visitor stops at a frog tank, our system would mistakenly model the visitor’s interest of ‘frog’ as a Frenchman rather than an amphibian. One language technology tool that allows us to achieve such disambiguation of word senses is SenseLearner, which is a minimally supervised sense tagger using the senses from WordNet (Mihalcea and Faruque 2004). It parses the text and tags surrounding each polysemous word, and combines the syntactic knowledge from the parser and the semantic knowledge from SemCor, in order to determine (learn) the exact sense of each polysemous word in a given context. We use the following example to illustrate the application of the system:

A sentence from a document on frogs:

Male frogs usually call from rocks on stream or river banks.

After disambiguation by SenseLearner:

male/JJ#1 frog/NN#1 usually/RB#1 call/VB#1 from/IN rock/NN#1 on/IN stream/NN#1 or/CC river/NN#1 bank/NN#1

Each word has been assigned a Part-of-Speech tag (e.g. JJ for *male* as an adjective). Words that have been assigned a ‘#’ symbol followed by an integer are polysemous words, where the integer is the *sense-ID*. In the example above, SENSE-LEARNER chooses the first sense (corresponding to ‘synset#1’ in Figure 3.2) to the instance of *frog* because the first synset provides the most appropriate interpretation in the given context. For this sentence, the content words of our interest are the ones with their POS tag and sense-ID highlighted in **bold**, i.e. nouns. We will show an application of SENSELEARNER later. (Section 4.2.4).
Chapter 4

Methodology

In our research, we design and experiment four different computational models, contrast their performances, and see which one is most optimal for our prototype system to dynamically detect the visitor’s interests as he tours the museum. Each model has a unique methodology for representing the exhibit content, and as we extract certain features (words and their senses) out of this content to infer the visitor’s interests, we would expect all four models to infer them differently. In this chapter, we begin with a description of how we extract content words from the document, followed by an outline of our weighting scheme applied to these content words, which then leads to the details of our computational models. At the end we design an evaluation method which benchmarks what each model thinks the visitor is interested in against what the visitor is actually interested in.

4.1 Content Manipulation

As we stated earlier in Section 3.1.1, all documents used in our research are online documents, most of which are collected from the Melbourne Museum website, and a few from Wikipedia. For the three galleries chosen in our research (Australian, Forest, and Prehistoric Life), we first crawled all webpages within the scope of these three galleries and filtered out URLs that are related to ‘e-news’, ‘mvmembers’, ‘privacy’ etc. that do not correspond to any exhibit or exhibit area. From the remaining pages, we filter out PDF and binary files for simplicity. The remaining pages are assumed to carry information about particular exhibits or topics. As we also mentioned in Section 3.1.1, not all documents are unambiguous in their correspondence with single exhibit(s). Within the scope of the three galleries, we found that roughly 30% of online documents are ambiguous. For instance, http://www.museum.vic.gov.au/dinosaurs/lifetime-oceans.html gives a description on at least 4 separate subtopics: ‘Trilobites’, ‘Fish and their Evolution’, ‘Amphibians and their Evolution’, and ‘Diversification in the Sea’. Each of these corresponds to a unique exhibit in
the ‘Prehistoric Life’ exhibit area. The solution to this is to segment the document into its sub-topics. In effect, this document is partitioned into four child documents, each of which corresponds to one exhibit. For each document, we tokenise the text, remove stopwords, and run the remaining words through a stemmer, as described earlier. After this linguistic processing of each document, we extract its semantic content (i.e. nouns) and add them to our vocabulary. After processing all documents, the vocabulary contains all distinct nouns across three galleries. We represent this vocabulary as an array of nouns, as shown schematically below:

(fossil, example, land, DNA, north, mammoth, cricket...)

There are close to 10,000 nouns in our vocabulary, drawn from nearly 700 documents. As is apparent in the above example vector, some nouns are less specialised in meaning than others.

### 4.1.1 Term Weighting

As we outlined in Section 3.2.2, nouns are the only semantic words that we are interested in, since they are more indicative of the exhibit content (hence the term ‘content word’); they are more likely to exist in multiple senses; and they make up the majority of words in our text collection. However, not all nouns are equally informative. For example, in a document describing ‘Ice Age Animals in Siberia’ that contains mammoth and amount among its nouns, how can we determine if one is indicative of the document content than the other? To determine this, we need a mechanism to associate each of these nouns with a weight to measure its individual value with respect to the exhibit content. Tf.idf makes up one component in our weighting scheme. It calculates values for each content word in a document through an inverse proportion of the word frequency in the document to the percentage of documents the word appears in (Strzalkowski et al. 1998). The tf.idf weight for each content word \( t \) in a document \( q \) is calculated as follows:

\[
tf.idf = \log_e(1 + f_{q,t}) \times \log_e(1 + N/f_t)
\]

\( f_{q,t} \): frequency of the term \( t \) in document \( q \)
\( f_t \): total number of documents in the collection that contain the term \( t \)
\( N \): total number of documents in the collection

Using this formula, content words with high \( tf.idf \) values relative to a given document imply a strong relationship with that document, i.e. words are frequent in this document but rare in the collection, so they carry specific meaning for this document only. In the above document on Ice Age Animals, the noun mammoth carries a \( tf.idf \)

weight of 4.32; whereas the word *land* carries a weight of 1.92. Hence we can say the content word *mammoth* contributes more to the content of this exhibit than the word *land*. By observation, this also makes sense because the word *land* occurs in a range of documents across our three galleries, and hence if the visitor stops at this exhibit, he is most likely to be interested in mammoths. For this particular exhibit, we can map its content words onto our vocabulary to create a document vector that shows the presence of words in the exhibit content, from which we then create a weighted document vector that indicates the importance of each content word, as shown below:

\[
\text{vocabulary: } \langle \text{fossil, example, land, DNA, north, mammoth, cricket...} \rangle
\]

\[
\text{document vector: } \langle 0, 0, 1, 0, 1, 1, 0... \rangle
\]

\[
\text{weighted vector: } \langle 0, 0, 1.92, 0, 5.08, 4.32, 0... \rangle
\]

Any content word in the vocabulary that is also found in the given document is marked with a 1 in the document vector, at the position where this content word locates in the vocabulary. At each of these positions, the \( \text{tf.idf} \) weight of the content word is assigned, creating a weighted vector for this document as shown above. For instance, the noun *land* may be at position 3 in the vocabulary, and so the 3\(^{rd}\) entry in the document vector is marked as 1; whereas the 3\(^{rd}\) entry in the weighted vector is the \( \text{tf.idf} \) weight for this word relative to the document.

Given a position in the vector, we can map it onto the vocabulary array to find the corresponding word. Hence, our weighted vector gives an indication of what content words exist in our document and how informative they are. Nevertheless, judging a content word based on its frequency information might not be reliable in all cases. By observation from the weighted vector above, the word *north* is given a higher \( \text{tf.idf} \) weight than *mammoth*, although in reality the visitor is not likely to be interested in this exhibit because it is about NORTH as a concept. As the \( \text{tf.idf} \) scheme neglects the information about the semantic content of each word, we need to embed this information somehow. This leads us into the second component in our overall weighting scheme.

### 4.1.2 Semantic Depth

Important words within the exhibit content can generally be useful to model a visitor’s areas of interest. This is the case for the word *mammoth* as we saw earlier. Besides using the rareness of a word, we also propose the use of semantic depth as another property of a word to indicate its importance. In our experiment, the semantic depth of a word (or a word sense) is a measure of how specialised it is in meaning. We investigate this degree of specialisation using the notion of a concept hierarchy, where \{noun: entity\} is chosen as the top-level hyponym (i.e. the most generalised
Figure 4.1: Hypernym Hierarchy for first synset of *mammoth*, depth = 14

concept). In order to work out the degree of specialisation for a word sense, we look at the path from its synset to the top-level hypernym, and calculate the distance of this ‘hypernym path’. Figure 4.1.1 shows a hypernym path from the top-level hypernym down to the synset that corresponds to the most frequent sense of the word *mammoth*.

This figure shows the hypernym path from the most general concept, namely, {noun: entity}, to the more specialised concept of *mammoth* (hyponym). The path distance is 14, i.e. the total number of concepts along this path, and hence 14 is the semantic depth for the synset {noun: mammoth}. This way, for a given word sense, we can use its synset to determine how specialised it is relative to its distance from the root node. For a given word, on the other hand, we take the average across the semantic depths of all its synsets. In our experiment, if a given synset is observed to be specialised in meaning, then we assume its corresponding word (or word sense) is important for the modelling of an exhibit content. For the above hypernym path, we can say that *mammoth* is specialised in meaning and is an important concept for the whole exhibit. This then allows us to infer concepts of the exhibit that could be of the visitor’s interests.

Our overall weighting scheme combines the *tf.idf* scheme with the calculation of semantic depth. To achieve a balance between both components, we take the $\log_e$ of the semantic depth value as shown below:

$$W(X) = \log_e(\text{semdepth}(X)) \times \text{tfidf}(\text{word}(X)) \quad (4.2)$$

$X$: a content word or a word sense

$\text{semdepth}(X)$: semantic depth of the synset corresponding to $X$

$\text{word}(X)$: the word that $X$ belongs to, if $X$ is a word sense

Using this scheme, the weight given to the word *north* is 7.34, whereas *mammoth*
is assigned a weight of 11.25, which now makes more sense since *mammoth* is more closely related to the semantic content of the exhibit. In Section 4.2, we will see that the differences between our computational models depend on the nature of $X$ – whether it is a content word or a word sense, and if is a word sense, how it is disambiguated.

### 4.1.3 Word Sense Disambiguation

For the noun *cricket* in our vocabulary, one cannot tell exactly what it refers to. Polysemous words such as this have multiple senses and carry different meanings depending on the context (topic). This is why we need to use context information to distinguish between senses. For the purpose of modelling a visitor, we need to be cautious when using polysemous nouns to represent exhibit content. For example, if the visitor is in the Forest Gallery and is currently looking at an insect exhibit displaying a cricket, then we should detect that this cricket refers to a leaping insect; whereas if the visitor is touring through the Australian Gallery and stops at the photo showing the Victorian Cricket Club, then we should detect that this visitor is interested in cricket as a sport. Therefore, the same word can represent completely different concepts in the museum, and hence can represent completely different interests of a given visitor. If we simply assume that every content word has only one sense in all contexts, then one would predict low accuracy in learning about the visitor and thus deliver potentially irrelevant content to him.

This provides the motivation for *Word Sense Disambiguation* (WSD). WSD is a language processing technique that, from all possible senses of a word, selects the most probable sense (Yarowsky 1992), given the linguistic or real-world context. This forms the basis of the linguistic models in our research, where we experiment with different mechanisms to decide on which sense of a given polysemous word is the most appropriate in a given context. This decision is crucial in disambiguating between the visitor’s actual interests and potential noise introduced by polysemous words. This is why we want to rely on the word senses rather than the words themselves to model the semantic information about an exhibit. We explain why this is the case with regard to our four computational models.

### 4.2 Computational Models

The overall goal for the design of the following four different computational models is to show that it is possible to abstract away from words (which can be ambiguous, as shown above) to semantics so to be more context-sensitive and informative in representing the exhibit content, and hence the visitor. The prototype system will then adopt each model at a time to test its accuracy in modelling the visitor’s interests.
Chapter 4: Methodology

We expect this accuracy to increase as a result of our transition from using words to semantics.

### 4.2.1 Word-Level Model

The word-level model provides the most basic way of representing data in the collection, since it only deals with words and does not take into account the senses of polysemous words in different contexts.

In this model, we focus on assigning weights to content words, i.e. the \( X \) component in our weighting scheme (Equation 4.2) is the content word. Hence, each document (exhibit) is represented as a collection of content words. The algorithm to model all documents is as follows:

/* Stage 1: forming the vocabulary using words */

```plaintext
VOCAB = {}
for DOC in DOC_COLLECTION do
  identify all unique content words in DOC
  insert words to VOCAB
end for
```

/* Stage 2: allocating weights to word-level entries */

```plaintext
for DOC in DOC_COLLECTION do
  initialise DOCVECTOR of length LEN(VOCAB)
  identify all unique content words in DOC
  for each content word W do
    calculate TFIDF(W)
    calculate AVERAGE_SEMANTIC_DEPTH(W) across all word senses
    T := INDEX(W) in VOCAB
    DOCVECTOR[T] := TFIDF(W) \times \log_e(AVERAGE_SEMANTIC_DEPTH(W))
  end for
end for
```

The above algorithm is only designed for our word-level model. At the end of this algorithm, each exhibit would correspond to a weighted document vector of the same format as the weighted vector shown in 4.1.1. We now provide an algorithm to show how these vectors are used to model the interests of a visitor:

/* Stage 3: using weighted vectors to model the visitor */

```plaintext
initialise USERVECTOR of length LEN(VOCAB)
for EXHIBIT in USER_PATH do
  find the weighted DOCVECTOR for EXHIBIT
  for i = 0 to LEN(VOCAB) do
    USERVECTOR[i] += DOCVECTOR[i] \times \log_e(TIME_AT(EXHIBIT))
  end for
end for
```
normalise USER\_VECTOR

end for

Stage 3 shows the algorithm for modelling the user model, and is applied in each of our computational models. The normalisation of vectors is an essential in computing cosine similarity on vectors (see Section 4.3). To explain the above algorithm mathematically, we use a vector (of the same length as each weighted document vector) as a representation of the visitor’s profile. The $i^{th}$ entry in this profile vector is represented as $u_i$, where:

$$u_i = \sum_{d \in D} W_{i,d} \times \log_e(t_d) \quad (4.3)$$

$D$: list of exhibits visited
d: each exhibit as a document vector
t: time (seconds) spent at a given exhibit, $t_d > 0$
$W_{i,d}$: weight at position $i$ of document vector $d$

As we said earlier, the history of visited exhibits and the time spent on each exhibit both provide an indication of the visitor’s context, i.e. the visitor’s state in relation to the physical space. This in turn can be used to infer the visitor’s interests at a particular point in time, which would accumulate to form a user profile by the end. For the time factor, $t_d$, it is not intuitive to assume that if a visitor spends twice as long at an exhibit then he is twice as interested in it. The use of a logarithmic scale is the standard approach for this type of scenario, and hence we use it on $t_d$ here.

Further, the above notation looks at the $i^{th}$ entry of each document vector and computes the sum across all document-vectors. What this means is, for instance, if the $i^{th}$ entry corresponds to the word *dinosaur* in our vocabulary, and both vector $\vec{A}$ and $\vec{B}$ (corresponding to exhibits A and B, say) have weights 0.5 and 0.6 at their $i^{th}$ entry respectively, then once the visitor has seen exhibits A and B, we can be confident that he is likely to be interested in dinosaurs and hence we increase the weight at $i^{th}$ in the visitor’s profile by 1.1 (i.e. 0.5+0.6). Hence, this concept stands as a representation of the visitor’s interests. As the visitor sees more exhibits related to dinosaurs, the weight of $i^{th}$ entry will increment in the visitor’s profile. This is how we update our vector representation of the visitor as he tours the museum.

The basic steps of the above algorithm are used in each model. However, what differs is the data type of each $i^{th}$ entry in our vocabulary. In this word-level model, our vocabulary is made up of words, and each entry in the vector is a numeric weight representing the importance of a word relative to document. In other models, the vocabulary is made up of word senses, as we will see shortly. The word-level model works fine for monosemous words such as *dinosaur* because each of these words only refers to a single concept regardless of the context. Nevertheless, for words such
as cricket, our word-level model neglects the fact that one document might refer to cricket as an insect, whereas another document relates cricket to sports. Therefore, using the syntax of words to model the visitor’s interests is deemed to result in some misjudgment, and hence we do not expect this model to accurately learn the visitor’s interests. This will be discussed further in Chapter 5. Now, let’s consider several alternative models that not only look at the words, but also their semantic meaning.

### 4.2.2 Uniform tf.idf Weight Distribution

Unlike the word-level model, the ‘Uniform tf.idf Weight Distribution’ (UTWD) model takes the sense(s) of each word into consideration. Instead of relying on the syntax of words in the document, we use WordNet to retrieve the synsets of each word and use them to create our vocabulary. As we just saw, the data type of each entry in VOCAB, DOC-VECTOR and USER-VECTOR shown in the previous algorithms is taken to be a word in the word-level model. In all three other models (UTWD, MFS and LS), this data type is taken to be a synset. Hence, the vectors in these models are significantly longer and more complex than that of word-level model. This means we cannot directly compare the vectors from UTWD against vectors in word-level model. We will refer back to this point in Section 4.3.

Each computational model has a unique method to assign numeric weights to entries in each vector. For the UTWD model, the method is as follows:

```c
/* Stage 1: forming the vocabulary using synsets */
VOCAB = {} 
for DOC in DOC_COLLECTION do 
    identify all content words 
    for each content word W do 
        identify all synsets of W 
        insert all synsets to VOCAB 
    end for 
end for 

/* Stage 2: allocating weights to synset-level entries */
for DOC in DOC_COLLECTION do 
    initialise DOC VECTOR of length LEN(VOCAB) 
    identify all content words in DOC 
    for each content word W do 
        identify all synsets for W 
        UNIFORM_TFIDF := TFIDF(W)/COUNT(synsets of W) 
        for each synset S do 
            calculate SEMANTIC_DEPTH(S) 
            T := INDEX(S) in VOCAB 
        end for 
    end for 
end for 
```
Chapter 4: Methodology

\[ \text{DOC VECTOR}[T] := \text{UNIFORM}_\text{TFIDF} \times \log_e(\text{SEMANTIC DEPTH}(S)) \]

The algorithm for using weighted vectors to model the visitor is the same algorithm as in the word-level model (Stage 3), except in this case, the data type of each vector entry is a synset. The reason we utilise the synsets of a given content word is because each synset corresponds to a particular sense of the word, and in turn gives information (synonyms) about each word sense. As each content word has its \text{tf.idf} weighting, we assign an equal share of the weighting to each sense of the word. The reason for this is based on the naive assumption that if a word \( W \) appears in a given document, then each sense of \( W \) has equal likelihood to occur in the context of this document. Considering our earlier example of the polysemous word \( \text{frog} \), the UTWD model treats each of the three word senses as equally probable in the museum context, meaning that if the visitor is seeing a frog, then this ‘frog’ can refer to the classes of \text{FRENCHMAN}, \text{ADORNMENT}, or \text{AMPHIBIAN ANIMAL}. The model does not know which concept to pick to model the visitor, and hence it assumes the visitor is interested in all three of them. Clearly, this does not make sense, since it disregard the context for the word \( \text{frog} \), which we know is related to the class \text{AMPHIBIAN ANIMAL} (at least within the scope of the three galleries).

It is worthwhile reminding ourselves at this point that our hope is to design a representation for each exhibit content where we abstract away from words to synsets. The UTWD model is the first step in our transition from using content words to their synsets, but it remains clear that we still need to make a decision on the choice of a synset, rather than taking all synsets into consideration and blindly assigning weights to synsets that do not match the actual context of the exhibit.

4.2.3 Most Frequent Sense

As we said above, for each polysemous word in the document, we do not expect it to occur in each of its senses with equal probability, and hence we need to explicitly choose a word sense. This means we need to choose one synset for each occurrence of a given word. One simple way to choose a synset would be to select, out of all senses for a given word, the sense that is most frequently used. This is the method adopted by our third model, the ‘Most-Frequent-Sense’ (MFS) model, which makes a decision on the choice of synset based on how frequent the corresponding word sense is used. The frequency of each sense is pre-determined using WordNet, which uses the SemCor corpus as the training set. By convention, for each word, WordNet organises the synsets in a way such that the first synset corresponds to the most frequent word sense. For example, in Figure 3.2, the first synset is related to the context of ‘amphibian’, which means the most frequent sense for the word \( \text{frog} \) is related to an amphibian.

Unlike the previous UTWD model, our MFS model does not uniformly split the \text{tf.idf}
weight across all senses of each word. Instead, it assigns the entire weight to the most frequent word sense. With respect to the algorithm for MFS model, both stage 1 and 3 are the same as in the UTWD model, and only stage 2 differs:

/* Stage 2: allocating weights to synset-level entries */
for DOC in DOC COLLECTION do
    initialise DOC VECTOR of length LEN(VOCAB)
    identify all content words in DOC
    for each content word W do
        calculate TFIDF(W)
        S := first synset of word W stored on WordNet
        calculate SEMANTIC DEPTH(S)
        T := INDEX(S) in VOCAB
        DOC VECTOR[T] += TFIDF(W) × log_e(SEMANTIC DEPTH(S))
    end for
end for

The motivation for the design of our MFS model comes from the observation that, for each polysemous content word, its most frequently-occurring sense (according to SemCor) often happens to be the sense that fits with the context of the museum. This is the case for words such as spider, china, frog, etc. that all consistently exist in their most frequent sense in the museum documents. We can safely use the first senses of these words to model interests of our visitor. For words such as cricket, however, it has two senses: the first (most frequent) sense refers to an insect, the second refers to a sport, and both senses exist equally likely in the document collection. Hence, if we choose the first sense every time we see the word cricket, we can correctly model the visitor in the context of the forest insects, but we will be guaranteed to misjudge his interest within the context of Victorian sport. Despite this, we hypothesise that the MFS model will outperform both the word-level and the UTWD model since the content words within museum documents generally exist in their first senses anyway. The problem is, our selection of the word sense is fixed regardless of the actual context. Therefore, more flexibility is needed to detect the visitor’s context and choose the right word sense based on this context. Let us see how this flexibility can be computed into a model.

4.2.4 Learnt-Sense

‘Learnt-Sense’ (LS) is the last computational model investigated in this research. It uses SenseLearner as a preprocessor, aiming to incorporate the sense-sensitivity that MFS failed to address. As we said earlier, SenseLearner is capable of predicting which word sense is applied for a word in a given context, and hence we hope to use this tool for the purpose of Word Sense Disambiguation (Section 4.1.3). In our previous example in Section 4.1.3, we showed how SenseLearner labels each polysemous word with the predicted sense-ID. Take the word frog for instance, SenseLearner is able to select its
first sense \((\text{frog}/\text{NN} \#1)\) as the most probable choice in the given context. Therefore, if the user has visited the frog tank, we will use the first synset of the word \textit{frog} to indicate that this user might be interested in frogs as amphibians.

The algorithm for this model does not differ from the MFS model except the way in which the weights are allocated: for each word, the MFS model uses WordNet to retrieve the most frequent sense of this word, and assigns the weight to the synset corresponding to this sense; on the other hand, the LS model uses \textsc{SenseLearner} to predict the sense of a given word, and then assigns the weight to its corresponding synset in the vocabulary. The details of calculations and vocabulary formulation (consisting of synsets not words) are the same as in the MFS model.

We hope \textsc{SenseLearner} can deliver high accuracy in context-sensitive sense predictions. Mihalcea and Faruque (2004) explain how \textsc{SenseLearner} uses semantic knowledge from WordNet and syntactic knowledge from parsed text to intelligently learn about the context. Hence for each polysemous word, \textsc{SenseLearner} is capable of selecting the sense that is most appropriate for the given context. Also, \textsc{SenseLearner} is able to disambiguate a content word even if it did not appear in the training data (Mihalcea and Faruque 2004). Due to these reasons, our Learnt-Sense model is expected to outperform all previous models. The model predicts the word sense, and assigns weights to the corresponding synset to represent the information of each exhibit. These weights can then allow us to judge which parts of the exhibit content are useful for inferring the visitor’s interests.

4.2.5 Summary

In summary, we use the word-level model to represent an exhibit content as a collection of content words. We then design the UTWD model to abstract this representation from a collection of words to their synsets, treating the synsets as if their corresponding word senses are equally probable. To improve on this, we use WordNet to choose the most frequent sense for each content word, hoping the MFS model fits well with most of the museum documents. What really adapts to the museum context, however, is expected to be the LS model, which selects word senses based on the exhibit context. Each model can be seen as being more computationally sophisticated than the previous, since additional linguistic knowledge of the text is computed at each step of our abstraction from words to synsets. In the following section, we will describe our method to evaluate each of these models, based on the deviation between the inferred interests of the visitor and his actual interests.

4.3 Experimentation and Evaluation

Each of the four methods described above attempts to model the visitor differently, giving us a vector representation of what the visitor \textbf{might} be interested in. In our research, we need to compare each of these against our gold standard – what the visitor \textbf{is indeed} interested in. To build such gold standard, we harvest direct feedback from the visitor,
Table 4.1: A fragment of the Feedback Grid: Tell us your real interests.

<table>
<thead>
<tr>
<th>dinosaur</th>
<th>fern</th>
<th>ecosystem</th>
</tr>
</thead>
<tbody>
<tr>
<td>ant</td>
<td>gargoyles</td>
<td>reptile</td>
</tr>
<tr>
<td>racehorse</td>
<td>water</td>
<td>Olympics</td>
</tr>
<tr>
<td>glider</td>
<td>DNA</td>
<td>trilobite</td>
</tr>
</tbody>
</table>

asking them to explicitly outline what their interests are, from all exhibits visited. This feedback mechanism is only used at the end of the tour. In Table 4.1, we show a fragment of the ‘feedback grid’ that allows the visitor to key in their interests.

In reality, this ‘feedback grid’ is an interactive screen showing a table of words, allowing the visitor to press each word that reflects their interest(s) in the tour of the three galleries. We refer to each of these words in the grid as a selectable, and each selectable is also an entry in our word vocabulary. Each selectable is either monosemous, i.e. has a single word sense and is therefore unambiguous in any context (e.g. dinosaur); or in-context monosemous, i.e. has multiple senses, but we can easily and uniquely determine which sense is relevant to the museum context (e.g. frog). These words are chosen to make up the grid for two reasons: firstly, for the visitor’s convenience: the visitor does not need to disambiguate between possible senses of a word before choosing it; secondly, for our convenience: by looking at these words, we can uniquely predict their most appropriate senses in the museum context as our gold standards. This means that, for a word such as frog which has three synsets in the vocabulary, we only recognise its first synset, which we can predict to be the only synset relevant to the museum context. In effect, we are treating these words as if they only have one sense within the museum context, hence the term in-context monosemous. For these reasons, our table of selectables does not include words such as cricket and bat, because we cannot pre-determine their meanings without considering their specific contexts.

The feedback from the visitor contains a number of selectables, which then allows us to build our gold standard vector and compare it with the vectors from our models. This comparison is made using cosine similarity as shown below. It is a standard method in the research of information retrieval for comparing vectors based on their angle of deviation, while reducing the effect of vector length on the similarity score:

$$\text{cosim}(\vec{k}, \vec{g}) = \frac{\vec{k} \cdot \vec{g}}{||\vec{k}|| \cdot ||\vec{g}||}$$ (4.4)

$\vec{k}$ : resultant vector of each model to represent the perceived visitor’s interests
$\vec{g}$ : the gold standard vector of the visitor’s actual interests, from his own feedback
$||\vec{g}||$: the norm(length) of vector

If the visitor only presses on the word frog, say, we assign the value of 1.0 to the first synset of frog, and 0 to the other two synsets, in our gold standard vector. From our
inferred user vector, we then see if the first sense of *frog* is assigned a weight. If so, this weight will then be aligned (using cosine similarity) with the value 1.0 within our gold standard, indicating that we have inferred this visitor’s interest correctly.

The detailed algorithm is shown below. It follows on from stage 3 of our previous algorithm, in order to show the flow of methodology.

```plaintext
/* Stage 4: forming the gold standard */
/* USER_VECTOR has ALL the interests we inferred from visitor */
/* obtain USER_VECTOR learnt by a given model (from Stage 3) */

/* GOLD_STD contains entries corresponding to what visitor keys in */
/* initialise GOLD_STD of length LEN(VOCAB) */

for each selectable word B chosen by user from the grid do
  if using Word-level model then
    T := INDEX[B] in VOCAB
    GOLD_STD[T] := 1.0
    continue monosemous(B)
  S := SYNSET(B)
  T := INDEX(S) in VOCAB
  GOLD_STD[T] := 1.0
  else
    /* must be in-context monosemous then */
    S := PREDICT_SYNSET(B)
    T := INDEX(S) in VOCAB
    GOLD_STD[T] := 1.0
  end if
end for
/* Now, GOLD_STD = list of visitor’s REAL interests */
/* Stage 5: compare ‘inferred visitor’s interest’ vs. ‘actual’ */
/* initialise INFERRED_IST of length LEN(VOCAB) */
for each offset F in GOLD_STD assigned with 1.0 from Stage 4 do
  INFERRED_IST[F] := USER_VECTOR[F]
end for
/* Now, INFERRED_IST = list inferred visitor’s interests */
/* normalise, and assimilate our inferred against gold standard */
SCORE = cosim(INFERRED_IST, GOLD_STD)
```

The resultant score is within the interval [0, 1]. This gives us recall as our performance measure, i.e. out of all correct interests for a specific visitor, how many interests did each of our models pick up. We do not measure the precision (i.e. out of what we think the visitor is interested in, how many is he really interested in) because this depends directly on the number of non-zero entries in the USER_VECTOR. In the case of museum documents,
more than 50% of such entries correspond to ‘noise-words’ such as *north*, *amount* etc. that are not likely to help us to model the visitor’s interests. The precision would be based on words of this nature, and hence corrupt our measure of how *real* interests are detected and modelled. Recall, on the other hand, is measured based on these real interests, and hence the effect of ‘noise-words’ is significantly reduced.

In summary, this chapter has detailed the design motivation of each computational model in our research. The sketch of algorithm in a flow of five stages is the core of our methodology. The following chapter shows the results we obtained and their analysis.
Chapter 5

Results and Discussion

Table 5.1 compares the effectiveness of the four computational linguistic models. The experiment was performed over five individual users, each of whom approached the task differently. Each score in the table gives an estimate of how well the inferred visitor’s interests match up with his actual interests under a given model, in the form of recall. The recall is within the interval \([0, 1]\), where high recall in this case corresponds to high accuracy in the detection and modelling of the visitor’s interests. We now analyse the performance of each model individually.

5.1 Word-Level

We hypothesised in Section 4.2.1 that the word-level model would not yield accurate learning of the visitor, and indeed, its performance is ranked second last. Since the content of vectors used in this model completely ignores sense and synset information, it is not surprising that it results in poor detection of the visitor’s actual interests (consider the cricket (insect) vs. cricket (sport) scenario again). As we outlined earlier, our hope in this research is to abstract away from words to semantics. Yet, the method for user-modelling adopted in our word-level model relies on a syntactic level of information processing, and hence it has a low degree of content understanding and is a bad choice to model objects in an information-rich environment. The fact that this model achieved the best result (0.493) across all four models does not indicate its superiority in any way. As the number of polysemous words increases, the Word-level model is the only model that does not suffer, since its vocabulary assumes all words as if they have only one sense. This is the worst assumption to make with respect to WSD, which is one of the main themes of our research. The seemingly reasonable results obtained by this model are all based on this assumption.

5.2 UTWD: Uniform-tf.idf-Weight-Distribution

The UTWD model yields the worst performance out of the three models that take synsets into account. This model assumes that all senses of a given polysemous word have the same probability distribution in any context within the museum. This has clearly been
proven as a naive assumption, given that a significant number of senses cannot be treated as equally probable (consider the three synsets for frog again). Under this model, improbable senses all obtain an equal share of the tf.idf weight, which essentially contaminates our overall purpose to ‘identify one most suitable sense for each word and use it to model one possible interest of the visitor’. As a result, the actual probable senses do not get a fair share of the tf.idf weight, hence the UTWD model yields the worst performance in all categories. This model marks the beginning of abstraction from words to synsets, one would hope we can do better as more linguistic features are considered in our model.

5.3 MFS: Most-Frequent-Sense

On average, the MFS model yields the most accurate modelling of the visitor’s interests. The assumption behind MFS was that the ‘the most frequent sense of any word in the museum document is the sense that fits with the exact context of the document’. For most content words, as they are in-context monosemous (i.e. their sense can be easily and uniquely predicted) with respect to the museum context, this model works fine. However, as we saw earlier, for highly polysemous words such as cricket, its most frequent sense refers to the ‘leaping insect’ which is clearly inappropriate in the context of Australian sport. Although our MFS model often mis-predicts the correct sense in these situations, such situations are only the minority in the documents across our three galleries. We hope SenseLearner will be able to deal with these highly ambiguous words by being context-sensitive in the selection of word senses.

5.4 LS: Learnt-Sense

On average, the LS model did not perform as well as expected. In our description of the LS model earlier, we hypothesised that this model will produce the most accurate user model of all, since it uses SenseLearner as an external toolkit to select the most probable sense for each word in a given context. However, according to Mihalcea and Faruque (2004),

<table>
<thead>
<tr>
<th>Models</th>
<th>User#1</th>
<th>User#2</th>
<th>User#3</th>
<th>User#4</th>
<th>User#5</th>
<th>Avg on 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO.terms selected</td>
<td>5</td>
<td>10</td>
<td>2</td>
<td>6</td>
<td>16</td>
<td>7.8</td>
</tr>
<tr>
<td>NO.polysemous</td>
<td>2</td>
<td>6</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>2.4</td>
</tr>
<tr>
<td>UTWD</td>
<td>0.129</td>
<td>0.076</td>
<td>0.0</td>
<td>0.117</td>
<td>0.146</td>
<td>0.094</td>
</tr>
<tr>
<td>MFS</td>
<td>0.609</td>
<td>0.435</td>
<td>0.0</td>
<td>0.694</td>
<td>0.731</td>
<td>0.494</td>
</tr>
<tr>
<td>LS</td>
<td>0.539</td>
<td>0.402</td>
<td>0.0</td>
<td>0.644</td>
<td>0.682</td>
<td>0.453</td>
</tr>
</tbody>
</table>

Table 5.1: Assimilation Score: how well have we modelled the visitor’s interests? (Best results in each column in bold)
SenseLearner makes accurate predictions less than 70% of the time. Knowing this, we are then not surprised when SenseLearner mistakenly chooses the first sense for cricket (insect) when the sentences such as “...cricket game played in...” are in the documents. Furthermore, as SenseLearner was minimally trained on the SemCor corpus, the style of language in this corpus can be expected to differ from that adopted in the museum text. For polysemous nouns such as water, SenseLearner almost always select its first sense (i.e. \( H_2O \)); whereas in actual fact, the third sense (i.e. a water supply) is quite often used in the museum documents. Hence, cases such as this are all missed by SenseLearner. Having said this, however, in some situations where MFS mis-judges the sense, the LS model tends to give the correct sense prediction, and vice versa. This is the case for words such as life and history, where the most frequent sense of each word may or may not fit with the exhibit context. Thus, it might be possible in the future to combine MFS and LS to explore their complementariness, and hence hope to give better performance on our prototype system.

5.5 Other Findings:

Although having more users participating in our experiments would help to increase the confidence level of our findings, we nevertheless observed some general patterns in our results that are beyond the analysis of each individual model.

5.5.1 Selectables vs. Accuracy

When giving feedback, as the visitor selects more terms of interests from the grid, the assimilation value tends to increase, meaning that we have modelled the visitor more accurately. The explanation for this is: as the number of the visitor’s (selected) actual interests increases, so does the probability of finding a matched interest from our USER VECTOR that has inferred interests. For example, if the visitor selects all of the words from the grid, then the accuracy score will be guaranteed to be better than if he chooses one word or two words. In the case of choosing two words, all models appear to completely miss out on the two interests of the third visitor shown in Table 5.1, hence resulting in all zero values across four models. This tells us that we should not ask the visitor to enter too many or too few words from the grid, in order to provide a gold standard that is comparable to our inferred user-profile.

5.5.2 Polysemous vs. Accuracy

As more polysemous words get selected, the accuracy of our user-modelling decreases. This is true for UTWD, MFS and LS, since the vocabulary in each of these models is made up of word synsets. If the visitor chooses the word water, say, then both LS and MFS would select the first word sense in practice, whereas the actual sense is the third one. However, as the visitor chooses the word dinosaur, which only has one word sense, then both LS and MFS would assign the weighting to this word sense, and there is no room for mis-judgment. Hence, we can learn that as the visitor chooses more polysemous words from the feedback grid, it becomes more probable for these two models to select the incorrect word sense. This
explains why, as seen from Table 5.1, as the second visitor chooses six polysemous words out of ten (and it so happens that she chose the word *water* as one of them), we observed a dramatic decrease in the assimilation score between the inferred and the real interests of the visitor. It is apparent that we need a better word-sense predictor to assist in the learning of context in this case.
Chapter 6

Conclusion and Future Work

6.1 Conclusion

In this research, we have designed, implemented, and contrasted various computational models to explore the detection and modelling of the visitors’ interests in a museum environment. In this information-rich environment, whether a visitor wants to see a particular exhibit depends on how well the exhibit content is represented to fit with his areas of interest. We can, in some scenarios, use word tokens such as dinosaur and frog from the exhibit’s document to represent a visitor’s interests. Nevertheless, this is infeasible for polysemous words that can represent completely different interests of the user, such as cricket as a sport and cricket as an insect. The solution is to incorporate word senses to be sensitive to the context of each exhibit. For this purpose, we developed four different computational linguistic models: Word-Level, Uniform-tf.idf-Weight-Distribution, Most-Frequent-Sense, and Learnt-Sense models. Each of them imposes a different degree of utilising word sense information to model the exhibit content, from which the visitor’s interests are then inferred.

In an experiment performed over five museum visitors, we find that the Most-Frequent-Sense model produced the highest degree of user-learning, given the fact that most words within our data collection have word-sense properties inherent to this model. We expected the Learnt-Sense model to outperform others but the inaccuracy of SENSELEARNER leads to incorrect predictions of word senses, failing to detect the visitor’s interest relative to a given context. The MFS model outperforms LS by 0.04 on average, but it is observed that both models can be complimentary in some cases, and thus one would expect their combination to give optimal user-modelling in our prototype system. Through the incorporation of linguistic knowledge such as word senses and synonyms with respect to the exhibit content, both models are discovered to yield more accurate detection and modelling of the visitor’s interests than simple Word-Level model. This proves to be consistent with our design goals for the models – to abstract away from words to semantics in our representation of the visitor’s interests.
6.2 Future Work

In our research, we have used synsets as individual entries of our vectors in three of our models. However, due to time constraints, we did not explore a wider application of synsets in user-modelling. For future work, first, we would consider using the actual synonyms in each of the synsets, e.g. using words such as toad, amphibian etc. for the first synset of frog. These synonyms can be used to link different words or phrases together if they refer to the same concept, e.g. the phrase George Bush has the same synset as President Bush. This gives us more knowledge to model the visitor (or user in general). Second, it is worthwhile in the future to exploit the role of Named Entity Recognition on user-modelling and context-understanding. We expect this method of linguistic analysis to offer more information about the exhibit content. Third, as the overall Kubadj Project aims to deliver personalised information to the visitor, it is also important to develop research on text summarisation in order to retrieve and present information passages to visitors in the museum. Last, it might be possible to incorporate the visitor’s feedback as he tours through the museum. This is when the system delivers information passages to the visitor and asks if they were relevant. Based on this feedback, the system can promptly adjust the learning of the visitor and deliver more relevant information.

Acknowledgment

I wish to thank Dr. Tim Baldwin and Dr. Steven Bird for their caring supervision and guidance throughout this memorable research experience; Karl Grieser for mentoring me through my research and providing much-needed data for this project; and everyone in the Kubadji Project for their sharing of ideas and inspirations.
Bibliography


Randell, Cliff, Erik Geelhoed, Alan Dix, and Henk Muller, 2006. Exploring the Effects of Target Location Size and Position System Accuracy on Location Based Applications.


Author/s: 
Yang, Michael

Title: 
Dynamically detecting and modelling the visitor’s interests in a museum environment

Date: 
2007-10

Citation: 

Publication Status: 
Unpublished

Persistent Link: 
http://hdl.handle.net/11343/39321

File Description: 
Dynamically detecting and modelling the visitor’s interests in a museum environment

Terms and Conditions: 
Terms and Conditions: Copyright in works deposited in Minerva Access is retained by the copyright owner. The work may not be altered without permission from the copyright owner. Readers may only download, print and save electronic copies of whole works for their own personal non-commercial use. Any use that exceeds these limits requires permission from the copyright owner. Attribution is essential when quoting or paraphrasing from these works.