# The Virtual Bookshelf Approach

Dr. Anders Broberg[1], Ulrika Hägglund[1]

[1] Umeå University, Department of Computing Science,
90187 Umeå, Sweden
{bopspe, ulrika}@cs.umu.se
http://www.cs.umu.se/~bopspe/

## Keywords

Information object, Information space, Virtual bookshelf

## Abstract

The amount of information that people have to handle in their everyday life has increased and more is asked of their cognitive ability to select and assimilate the information that is important to them. One way to facilitate the cognitive work is to match information objects with each other and select/filter the ones that are relevant. To do this selection objects must be represented in such a way that comparison is possible. This paper presents the virtual bookshelf approach, which means that the user collects information that he/she finds interesting, and from this collection an information space is created aimed to be used for comparing and deciding the relevancy of new objects.

## 1 Introduction

The information society of today is characterized by high complexity, fast flows, and rapidly increasing amounts of information. Mobile phones and the Internet are two technical revolutions with a very high impact on society. Both technologies have changed the way we communicate and spread information.

Finding tools that facilitate the process of filtering information is the overall task that we focus on. In order to do this we need to study the use of this kind of tools and the experience from this kind of situation. This implies a need for: a stable theoretical platform; a technical platform; and techniques and algorithms, especially for modelling and matching the field of information interest. This paper describes ways to work with an information collection with the purpose to filter. A model to describe different types of objects in a way that makes matching between them possible is presented.

The next section is an overview of how information can be structured and reduced to a feature space where it is possible to decide if new objects are relevant with respect to a user's collected information. Section three describes the virtual bookshelf approach and presents a model of an information object. Section four gives some examples of applications that use this representation of information objects to match objects.

## 2 Working with Information Spaces

Most of the entities (persons, documents or artefacts in general) that we meet in our daily life have some sort of data and information associated to them; these "objects" can be viewed as information objects. To facilitate matching between information objects they must be structured or represented in an information space. This section presents a number of steps and alternatives in the process to create an information space aimed for matching information objects. The first step in the process is to model the information objects and then ordering them in an information space. To make the information space manageable it can be necessary to reduce the complexity. The last step in this process of creating a structured information space is to interpret or group the information objects into clusters.

### 2.1 Modelling and Ordering Information Space

Information objects can be described with a number of features. One common method to organise information object in a space is according to a number of predetermined features, i.e. a feature space. In general the complexity and dimensionality of the feature space can vary. What features that really are relevant depend on the kind of application, etc. In many applications there is a need for reducing the complexity of the feature space.

### 2.2 Reducing the Information Space

When the information is organised in a feature space, the next step is to reduce the space. Two approaches are presented, one based on predetermined features called category-based and the other based on algorithms that reduce the space by finding associative patterns and is called similarity-based reduction.

**Category-based Reduction**

The category-based reduction is using a limited number of predetermined categories with a number of well-defined features that best describe each category. The categories and its features must be decided in advance. It is necessary that it is based on a well-known ontology that determines which features that best define a category; compare with databases where it is important to have an established definition of the features that define each post.

The categorisation is made automatically by methods that analyse for example a text and find the category that best describes the object.

The use of predetermined categories can result in a static system because it is hard to change the categories once they have been determined. Another problem with predetermined alternatives is that they can result in a stereotyped description of an object. More features make the system more flexible and dynamic but also more difficult to handle.

**Similarity-based Reduction**

An algorithm that reduces the feature space is Latent Semantic Indexing (LSI). LSI automatically discovers latent relationships in document collections using the truncated singular value decomposition, SVD [Deerwester et al., 1990]. Dimension reduction is used to remove the noise and at the same time retain the most relevant structure in the association of terms and documents [Berry et al., 1995]. The dimension reduction gives a number of advantages: the representation becomes denser and two related documents can be near one another in the reduced space without sharing any terms. This can happen if the terms in each of the documents occur in other documents.

### 2.3 Interpreting the Information Space

To be able to analyse and classify how relevant an object is there is a need to interpret and group information object in the information space. This grouping can be done with a predetermined, well-known ontology (compare with a library classification system) or by the user, similar to how she organises her bookshelf. Both methods result in a number of known clusters that are named in advance. Another approach is to base the grouping on a cluster algorithm, for example K-means algorithm. The cluster algorithm interprets the information space and results in a number of anonymous clusters.

### 2.4 Matching Information Objects

The information space that the process above results in can be used to facilitate matching between information objects. This matching can be between objects in the model, but also between objects outside the model and objects in the model. Commonly, the classification or matching algorithms used in this step are based on measuring some sort of distance between the information objects. The methods for judging a new information object utilise knowledge about existing objects. For example the KNN-algorithm places the new object into the information space and judges it against a predefined number of nearest neighbours. Instead of a predefined size of the neighbour set, a threshold value for the distance measurement can be used to define the neighbour set. Another approach that also utilises knowledge about the existing objects is to measure the distance between the new object and the existing clusters (from the interpretation step).

## 3 The Virtual Bookshelf Approach

To find out what someone is interested in (and/or is expert on) is to reach inside someone's mind. With the lack of direct information technological methods that can help us we must find indirect ways of getting at a person's information interest. Because of the constantly ongoing interaction between an individual's internal resources and his/her external resources [Kaptelinin, 1996], a person's interests and knowledge is highly coupled to what the person has in the bookshelf, at the desk and/or in the bookmark list in the web browser. Information that someone has gathered is closely related or even similar with what is in one's thoughts. Articles, magazines, CDs, books, URLs, etc. that a person has collected can reflect someone's interest. Documents that a person has written, both private and work related texts, also reflect the information interest.

The main idea with the virtual bookshelf approach is to imitate how we organise our personal bookshelves to be able to decide how relevant and/or interesting a new object is to a user's interests. Each person has her own idea about how to structure a bookshelf and it may not agree with some predefined library classification system.

### 3.1 Working with Information Space in the Virtual Bookshelf Approach

The grouping of objects in an information space utilises the organisation in a user's bookshelf where each part in the bookshelf represents a dynamic field of interest. The dynamics in the way we manage our bookshelf makes the feature space complex, and the virtual bookshelf is represented with a complex feature space without any predetermined categories. When reducing the feature space automatic methods that take advantage of the latent structure in an information collection are used.

However, when the purpose is to match different kinds of information objects with each other it is important to have a model that describes an object in a general way. The next section describes a proposed model of a general information object.

### 3.2 Modelling Information Objects

Our model of an information object is composed of a core object together with information of different types related to it. A core object can be a person, a place or a physical object. Some examples are documents, a museum, a mug. Together with different kinds of information associated with the core object it forms the information object. To get a nuanced picture of the object it helps if the object is described with the help of a number of different components describing different parts of the information object. In this way the information object can be described with the components that best suit the method of application that is used in a particular situation. All components may not be necessary all the time.

**Facts about**

Facts about is one of the components that help to describe the core object; for a person facts about could be age, gender, income etc., for a document it could be type of document (pdf, word) and for a store facts about could be location, owner, open hours. Facts about are overall and objective data about the core object that are recorded in a systematic way and can be found in databases. The facts-about component is important in this model because it can quickly and rather easily give a picture of what kind of object it is. The picture may rely on assumptions and preconceived ideas but in some applications that is enough.

### Content

One of the most important components in this model is the (information) content. The content component is composed of information that is carried by the core object. The owner of the core object is responsible for the content and what is included in it. It is composed of data and information that the owner of the object has gathered and/or included in the content. If the core object is a document the content consists of information contained in the document such as text, pictures and graphs. For a store it could be information recorded and contained as a part of the business, for example the cash register.

### Meta-data

The meta-data component in our model is data about the content. Typically for meta-data is that they are designed for a special purpose and are data about data. The meta-data component has many similarities with facts about and can be viewed as a part of that component but we have chosen to separate them. The separation is done because facts about are data related to the core object and are not related to the content while meta-data is data about the content. Meta-data for a document can be language and subject area.

### Annotation

Another example of what can be included in the model of an information object is notes that have been attached to it by other objects. This component is called annotation and can work as a Post-it note, a help to remember things about a certain object or things to do when one comes near the object. The information can be private to the "notifier" or public to other objects. The annotations can be more temporary; when a note has been used it is often taken away. Facts about is in comparison more static. Annotations are subjective opinions about something or someone and describe how an object is associated to other objects.

### Context

The surrounding affects how someone acts and behaves. Places have the ability to remind us, and to set us into a mood. This meaning can be private or shared for a larger group or the whole society. Humans communicate asynchronously by leaving messages in the public space. Giving information a physical position in absolute or relative coordinates is an important part of this scenario that will have great impact of people's daily situation.

The context is an important component in the model because the context affects how someone acts. The main parts in this component are geographic location and time; both can give valuable information of the context that an object is in.

There is scope for information that is not characterised by the components described in this model but for the moment the ones listed above will do.

## 4 The Service Platform and Applications

Implementing applications that help us manage and overview situations with a high degree of complexity and facilitate our cognitive work is the final aim. This implies a need for a service platform that can prepare and work with an information space. The service platform that we have developed is composed of a number of servers, the most important ones are the Relevance Server (RS) that takes care of the relevance handling [Brändström, 2005], the Information Radar Server (IRS) that handles geographic data and the Virtual Bookshelf Server (VBS) [Johansson and Johansson, 2004] that together with clients manages the virtual bookshelf. These servers communicate with each other to serve the applications with appropriate data.

Some examples of applications that are interesting are described below. The idea with *Overview* is to capture, analyze and display how people act, move and where people are, i.e. to serve the users with an abstraction of the current situation in the form of maps. *K-Map* makes a personal visualization of the information landscape (located information objects). The visualization is based on a user's interests and how relevant information objects are. Abstract maps over the information landscape are constructed as overlays on ordinary geographical maps. *Information radar* (IR) scans the information landscape of the vicinity for relevant information objects. The discovered objects and their relevance to a user's information interest are reported [Johansson and Johansson, 2004].

## 5 Future Work

The virtual bookshelf approach will be tested with real users to see if it is a suitable approach. When a number of users have collected information to their bookshelf it is time to let the users test different applications based on the virtual bookshelf approach and in that way get the possibility to evaluate the usefulness of the virtual bookshelf approach. The next step will be to measure performance both from a cognitive and a technical perspective with the goal to improve and refine the service platform and applications so it can facilitate everyday life.

## References

[Berry et al., 1995] Michael W. Berry, Susan T. Dumais and Gavin W. O´Brien. Using Linear Algebra for Intelligent Information Retreival. *SIAM Review*, 37(4): 573-595.

[Brändström, 2005] Daniel Brändström. IRRS - A LSI-Based Relevance Server for the Virtual Bookshelf Service Platform. Master´s thesis, UMNAD 587/05 Umeå University, Department of Computing Science, 2005.

[Deerwester et al., 1990] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer and Richard Harshman. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6): 391-407.

[Johansson and Johansson, 2004] Niklas Johansson and Olov Johansson. The Information Radar System. Master Thesis, UMNAD 516/04 Umeå University, Department of Computing Science, 2004.

[Kaptelinin, 1996]Victor Kaptelinin. Activity Theory: Implications. in B. Nardi ed. *Context and Consciousness: Acitivity theory and Human-computer Interaction*, Massachusetts Institute of Technology, 1996, 103-116.