informationresearch

VOL. 15 NO. 3, SEPTEMBER, 2010

An associative index model for the results list based on Vannevar Bush's selection concept

Charles Cole, Charles-Antoine Julien and John E. Leide
School of Information Studies, McGill University, Montreal, Quebec, Canada

Abstract

Introduction. We define the results list problem in information search and suggest the *associative index model*, an ad-hoc, user-derived indexing solution based on Vannevar Bush's description of an associative indexing approach for his memex machine. We further define what selection means in indexing terms with reference to Charles Cutter's 3 objectives for the dictionary catalogue, particularly the selection objective.
Method. A case study utilizing a structured-interview method illustrates and tests the feasibility of the model by comparing it the AquaBrowser's word cloud, a popular word association and visualization information retrieval system.
Analysis. System-derived indexing schemes such as Aquabrowser's ‘word cloud' attempt to facilitate user processing of the results list by showing concept paths associated with the user's query. These externally-derived paths may not be helpful for an exploratory information search. The proposed model, which facilitates the user's information need processing, appears better suited for this type of search.
Results. The case study provides anecdotal evidence indicating the potential benefits of the model over externally-derived indexing systems.
Conclusion. Large-scale testing of the model is required to confirm the results.

CHANGE FONT

Introduction

A user conducting an information search using an Internet search engine, an online public access catalogue, or some other information retrievalsystem is often faced with a long and confusing results list, which may cause information overload in the user. We specifically define the case where the results list leads to information overload in the body of this article. As we will show, the information overload problem is especially acute for a user who does not know the answer s/he is expecting to find, a type of search called a topic, subject or exploratory information search (Meadow *et al.* 2007). The user conducting an exploratory information search does not have a frame of reference for recognizing the target answer in the results list.

In contrast to the exploratory search is a search where users know exactly what they are looking for, which we label as known item or known answer search. We start from the assumption that a user conducting a known item or known answer-type search, which can be '*uniquely specified by attribute values*', has a strong answer framework in mind at the start of the search and can therefore more easily process the results list than a user who is exploring a subject or topic area and does not know what type of answer s/he is looking for (Meadow *et al.* 2007: 273). The difference in user mindset between the exploratory and known item or answer search, we will show, is important enough to be considered a qualitative difference (i.e., fundamentally different) not simply a quantitative difference (i.e., a question of the user's degree of knowledge about the sought for answer).

In an exploratory search, the user is conducting what has been termed an ill-structured problem search, which '*typically require[s] additional knowledge from external sources in order to better understand the starting state...*' (Pirolli 2007: 20). For this type of search, a more holistic perspective is required than for a known item or known-answer information search.

For an exploratory search, humans '*berrypick*' pieces of information a bit at a time (Bates, 1989), going back and forth between a more informal type of searching, where they engage in any number of information-loaded activities (including thinking), and a more concentrated, formal type of searching where they actively utilize an information retrieval system. In this way, the information need that starts and motivates this user's information behaviour gradually and in an iterative fashion takes shape in the user's mind.

In this back and forth movement through the physical space, the user's mind is bombarded with a jumble of thoughts, some of which associate themselves with the user's gradually evolving perspective on the information problem. The mind may throw these solutions or thoughts aside, but they are still half there in memory, associated with some idea or experience. (The notion that we try on '...different solutions to see if they fit' before a solution becomes 'articulate in [our] consciousness', has made it to the op-ed pages of the New York Times (Brooks 2009: A25).)

In the larger scheme of things, apart from information retrieval problems, these synchronous thought associations we hold in memory when we are thinking about a problem serve the purpose of *'combin[ing] simultaneous ideas into more complex ideas'* (Anderson and Bower 1973: 22). Popper (1975) has called these new ideas about the problem '*tentative theories*'. But this is true of information retrieval as well. With a more solid conception of what it is he/she is looking for, something the user can put his or her finger on because the user now knows what s/he is looking for, the user can go back to the system to gather supporting evidence for the tentative theory. This is the way our minds are built; the way we naturally become informed and acquire new knowledge; it is, broadly speaking, the way we may think.

In an influential article entitled *As We May Think*, Vannevar Bush (1945) conceptualized an information retrieval machine, called the memex machine, designed to facilitate how humans naturally think when they interact with their information-rich environment, which he referred to as thinking by association. A stimulus in the environment, on the printed page, stimulates a thought that we for some reason associate with the stimulus. And it is this associated reaction that directs our next impulse to seek information. Bush's memex machine was designed to catch this associated reaction, allowing the memex user to instantaneously retrieve support information for this associated reaction, and the next one, and the next one, and so on, creating a trail through the information store that duplicated and supported the user's natural way of thinking about a topic.

Because Bush's machine was designed to facilitate the user's associative thinking by providing instantaneous informational support for it, the memex machine has been acknowledged as a model for the hypertext information environment of the Internet (Houston and Harmon 2007; Nelson 1991; Nyce and Kahn 1989); the Internet supplies instantaneous information for a hyperlinked concept that is embedded in the text the user is reading (Nelson 1991). (For the memex machine's influence on information science, see Cronin 2007; Smith 1991.)

Researchers have commented on the difficulty of creating an index scheme for the hypertext information search environment of the Internet, because this search environment is structured by the individual user (Hert, Jacob and Dawson 2000; Jacob 2004: 531; cf. also, Tebbutt 1999). Hert *et al.* (2000: 981) believe that such an index would involve an intermediate structure of some kind. By intermediate, we conjecture the structure would be between the user's cognition, the information retrieval system and the information store (see also Ingwersen and Järvelin 2005). But what form would such an indexing structure take?

General problem and conceptual framework

The general problem of this paper is to model an intermediate index structure for the hypertext information search environment on the Internet, with a specific focus on the results list. The model, called the Associative Index Model, is based on:

- the associative indexing concept which Bush himself briefly refers to in his article As We May Think, specifically his description of

associative thinking and the concept of selection; and
- a closer examination of selection as an information science concept through its seminal usage in Cutter's (1904) three objectives for a dictionary catalogue.

The *associative index model* collocates (i.e., brings together) the user's reactive thinking when perusing the results list, then facilitates the user's selection of the thought that best represents his/her information need at that particular moment. In this way, the model facilitates the user selecting the next item s/he will search for by either clicking on a citation hyperlink in the current results list or by formulating a new query.

In the final part of the article, we illustrate and evaluate in a case study the practicality of the model by having the study participant compare it to AquaBrowser's word cloud , a popular online catalogue results list visualization technology.

Bush's associative indexing concept

It is difficult now for us to understand the impact of Bush's (1945) conceptualization of his information storage and retrieval memex machine when he first described it in *The Atlantic Monthly*. It was only later, in a version of the article that appeared in *Life Magazine*, that drawings of the memex were added. The memex was represented as a microfilm reader-cum-computer, sitting on what looked like an office desk, underneath which was, stored on microfilm, all of human knowledge. A document text from the microfilm database appeared on a screen in front of the user. Any thought the user associated with the text being read on the screen could be instantly supported by the user clicking on the appropriate concept button at the bottom of the screen, which accessed information about the concept from the microfilm database of world knowledge under the desk; and so on and so on as the user followed his or her thought associations. (One can see the similarities to today's hypertext search environment on the Internet.) The domain expert user thus created information trails through the information store which others could subsequently follow.

Bush saw the information trail created by the domain expert memex user as an index system, a 'mem(ory-)ex'-based index system conceived in opposition to traditional hierarchical-based index systems (Buckland 1992). Bush's view of how the human memory system naturally directed human information seeking was therefore at the heart of the memex. However, his conception of an expert-created trail index system has been criticized (Buckland 1992). But his second suggestion, which he called 'associative indexing', appears to have greater potential as a model for indexing the Internet results list. Unfortunately, Bush said very little about this second indexing system, requiring us to extrapolate what it is from his description of his basic concepts of selection and associative thinking.

Associationism

In the 1930s and 1940s when Bush was conceptualizing the memex, the dominant paradigm for human thinking and reasoning research was associationism (other paradigms were Gestaltism, behaviourism, structuralism and functionalism) (Houston and Harmon 2007). Associationism has a long history in western philosophy, starting from Aristotle's essay *On Memory and Reminiscence* (Anderson and Bower 1973: 16) and it remains an influential theory today, for example in neural network research (Cao *et al.* 2004), in information science (Houston and Harmon 2007) and computer science (e.g., Google's wonder wheel and AquaBrowser's word cloud). The two principal constructs of associationism are that human memory is composed of ideas and that these ideas are connected in the human memory system in an associative network (Plotkin 2004). This results in the pivotal associationism notion that '*one idea will elicit another*' in the associative network that constitutes human memory (Anderson and Bower 1973: 24).

We illustrate the mechanisms of associationism using Anderson and Bower's (1973) phase-approach associative memory theory. In the first phase, an environmental probe sets in motion a **matching** process between the probe and memory, which establishes a 'correspondence between the current input or probe and some piece of the associative structure in memory' (Anderson and Bower 1973: 238). The matching process consists of a '*cue-dependent **probabilistic search***' of the '*associative network*' (Raaijmakers and Shiffrin 1981: 93). A probability-based matching system of this sort establishes a best match rather than a perfect match or yes/no matching system. Because a **best match** matching system allows for imperfections, it produces what is called a search set (Ratcliff 1978). The best-match search set is established by the initiating probe, after which a second mental process **identifies** how much of the matched associations in the person's memory structure '*is in fact useable for encoding the current input*' and whether '*unwarranted conceptualizations*' are occurring in the matching process that should be deleted (Anderson and Bower 1973: 237, 243-245). We note here that, according to this associative memory theory, the identifying process weeds out unusable and/or unwarranted associations from consideration. In the next sections we refer to these associationism concepts in bold in more detail.

Associationism and Bush's concept of selection

For his associative indexing concept, Bush incorporated the associationism concepts described above. He envisaged a new form of indexing; a more human, natural way. '[The human mind] *operates by association*', he wrote. '*With one item in its grasp, it snaps instantly to the next that is suggested by the association of thoughts ...*' (Bush 1945: 106). The central concept of the memex's associative indexing system and '*the essential feature of the memex*', is the concept of selection:

> [The memex] affords an immediate step, however, to associative indexing, the basic idea of which is the provision whereby any item may be caused at will to **select** immediately and automatically another. This is the essential feature of the memex (Bush 1945: 107) (emphasis added).

Bush operationalizes his concept of selection when he describes the procedure of selection for a fingerprint matching machine:

> This process is simple selection: it proceeds by examining in turn every one of a large **set of items** and by **picking out** those which have certain specified characteristics (Bush 1945: 106) (emphasis added).

Bush's term '*picking out*' can mean three different things according to the Oxford Dictionary of Current English (1985: 554):

a. '*Taking*': which denotes the physical part of selection when the user obtains the needed item. In information retrieval, obtaining occurs when the user either clicks on a hyperlink or by some other method (i.e., by typing in a new URL, by taking the item from the shelf, or by obtaining the item from another library by inter-library loan.)
b. '*Identifying or recognizing*': which denotes an earlier phase of selection than (a), where the person first must identify or recognize an item.
c. '*Distinguishing*': which denotes a further refinement or phase to the user's process of identifying or recognizing a needed item, i.e., the user identifies or recognizes the needed item by distinguishing the needed item from surrounding objects in the set.

With these dictionary definitions, we can refine Bush's two step process (from his above quote), giving us the following, elaborated definition of his concept of selection:

1. The user first examines a set of items,
2. 'identifying or recognizing' from the set those with 'certain specified characteristics', through the act of
    a. 'distinguishing' the needed item from the surrounding objects in the set. Distinguishing therefore implies the act of eliminating unusable and/or unwarranted associations from consideration (see the associative memory theory discussion in the previous section).
3. Finally, the user takes or obtains the selected item (off the library shelf or by some other physical act like clicking on a hyperlinked citation).

We examine part two of this elaborated definition of Bush's concept of selection in more detail in the next section.

The identification concept in the rapid selector machine

We further specify the identification concept in part two of our definition (from the previous section) by briefly situating Bush's memex machine in the context of the overall development of an information retrieval machine which culminated in the memex, called the rapid selector machine. The rapid selector machine was first conceptualized in Germany in 1927 by Emanuel Goldberg before development shifted to the US in the 1938-40 pre-war period with a team led by Bush (Buckland 1992). The rapid selector was in response to the explosion of data of all kinds and the inability of traditional, hierarchically-based cataloguing and indexing systems to handle these new, information retrieval tasks (Burke 1992). Bush, for example, proposed building the rapid selector for the Federal Bureau of Investigation to facilitate fingerprint matching (Burke and Buckland 1994).

The rapid selector machine worked on the principle of a quick and accurate information retrieval of a required microfilm document through a coding system, referred to as '*the associations*', each of which represented an important element of the document. The associations were punched into a stationary card in prescribed positions. When the user inserted the stationary coded card into the selector machine, it was then matched to all documents in the database through the coded associations. '*A perfect match between the codes on* [the microfilm document abstract frame] *and the codes on* [the stationary] *card triggered the selection circuit*' (Burke 1992: 150-151). Buckland (1992: 286) refers to the matching as '*the coincidence of a pattern on the microfilm matching the pattern on the search card*'. Because this was a perfect match retrieval system, each of the association codes took on enormous importance. In effect, each '[important or elemental] *datum had its own identity*' (Otlet 1934, quoted in Buckland 1992: 290).

The key point that each association code in the rapid selector's matching process had its own identity allows us to further specify our development of Bush's selection concept, to the following form:

1. The user first examines a set of items,
2. '*identifying or recognizing*' from the set those with '*certain specified characteristics*', through the act of
    a. giving identity to specified characteristics,
    b. then distinguishing the needed item, by these specified characteristics, from the surrounding objects in the set, which involves:
        i. keeping the items in the set that have the specified characteristics,
        ii. weeding out unusable and/or unwarranted associations from consideration.
3. The user takes or obtains the selected item by some physical action.

At this point, we frame the discussion of Bush's selection concept, with its associated concepts of set, identifying or recognizing and distinguishing, inside traditional indexing's usage of these terms.

Traditional indexing: selection, set, identification or recognition and distinguishing

We further define Bush's concept of selection by referring to how this term and its associated terms of the set and identification or recognition and distinguishing are used in traditional indexing and cataloguing as they were seminally articulated in Cutter's (1904: 12) three objectives for a dictionary catalogue:

1. To enable a person to find a book of which the author, title or the subject is known.
2. To show what a library has by a given author, on a given subject or in a given kind of literature (poetry, drama, fiction).
3. To assist in the choice of a work as to its edition (bibliographically), or as to its character (literary or topical).

The Cutter objectives have been commonly interpreted as, respectively:

1. the finding objective (for a book of which the author, title or subject is known),
2. the identifying objective (by collocation) and
3. the selection objective (from among items previously identified as pertinent) (Svenonius 2000: 14-15). (For the re-interpretation of Cutter's objectives over the years, see Svenonius 2000).

In the following analysis, we interpret the intention of the three objectives as being very different. The first objective is for a known item search (Baker and Lancaster 1991: 200), while the second and third objectives are for subject and topic searching respectively. This is a subject of debate when we attempt to operationalize the three objectives in terms of their effect on information search, because in a conceptual sense they are aimed at catalogue performance. This means that the objectives are concerned with facilitating or supporting user information search as a second order goal, their first order goal being the creation of an effective catalogue, or catalogue performance only (for a discussion of this point, see Lee *et al.* 2006). There is also another area of disagreement concerning the interpretation of Cutter's famous objectives.

The second disagreement concerning the interpretation of Cutter's objectives is how they operate together for certain kinds of searches, but for others less so or not at all. We are able to do this because we interpret the first objective narrowly, as a finding objective for an item that is known to the user. From this point of view, the first (finding) objective appears to be straightforward. The user has information about a particular item and wants to see if the item exists in the collection and, if so, from where, through the location code, it can be obtained. Most users search for a known item by title because a search for a known item by the subject is less efficient (Baker and Lancaster 1991: 274); it is inefficient to search for an item that is known to the user through the subject of the book because it is like a keyword search (i.e., leading to many citations in the result list the user has to go through to find the known item) (Lee *et al.* 2006). Nevertheless, subject known item searching is done and we make the assumption that this is the concern of Cutter's first objective. The advantage of this narrow interpretation of the first (finding) objective is that, for operationalizing it in terms of information search, it avoids overlap with the second (identifying) objective when the item is known by the user. A wide interpretation where the first objective is interpreted to include known subject searches (as opposed to known item searches) brings the first objective, from an operational point of view at least, within the purview of the second objective (i.e., showing the user what the catalogue has on the target subject so that the user can identify the sought after item).

The second (identification) and third (selection) objectives, on the other hand, are dealing with greater user uncertainty and involve deeper cognitive operations on the part of the user. For example, the second identification objective, when it is implemented in a catalogue, is intended to bring like-items together in one place to facilitate the user finding the needed item, but it requires the user to make comparisons not only between records for items that are quite similar, so s/he can pick out just the element that makes one item better than the others, but it also requires the user to manage and supervise this matching process through a detailed mental image of the needed item in working memory.

A search taxonomy based on the concept of known item search

We divide searching into known item and unknown item searching; the latter includes subject/topic and exploratory browsing searching (University of California Libraries... 2005; Wildemuth and O'Neill 1995). Other terms frequently used are purposive, semi-purposive and undirected searching (Bawden 1986). The dichotomous term for known item search, 'unknown item search', is rarely used in today's search-type taxonomies (Lee *et al.* 2006), even when known item search is included in the taxonomy (e.g., Meadow *et al.* 2007: 278). Here, unknown item searching is useful because it allows us to include all types of searches on a continuum. There are no hard and fast categories in information search. In place of the usual discrete search categories in analysing the functions of Cutter's finding, identifying and selecting objectives, the known-unknown item search continuum allows us to use the sliding concepts of uncertainty and identification (see below).

Searching for a known item and searching for an unknown item are two entirely different activities, involving different user cognitive processes, but it is necessary to nuance this statement, which is the purpose of this section. Broadly speaking, in a known item search by author or title, the user knows the item before the search commences, at least approximately and wants this specific item and no other. (There are variations of a known item search, ranging from a user who knows exactly what edition of the known item s/he wants, to a user who not only does not know the edition but does not even know the known item has more than one edition, or has forgotten important accessing information about the known item. See below for a discussion of these points.) In an unknown item, subject or topic search, on the other hand, users are less interested in the physical or virtual item in which the subject or topic information appears than if the information found there satisfies their information need.

In Figure 1, we diagram all types of information searches in the above statement on a known to unknown item continuum, starting from a known item search where the user has complete and accurate information about the known item, located at the far left-hand side of the figure, and ending with an unknown item, full-on exploratory search, located at the far right-hand side of the figure. The continuum, in turn, is made up of two sub-continuums, one each for known item searching and for unknown item searching. While the discussion that follows focuses on four distinct points on the continuum, the continuum is meant to include all types of searches, including those we do not discuss. For example, a user wishing to see what the library has by the author Margaret Atwood is conducting a search that is in-between ***known item search 2*** and ***unknown item search 1***. In addition, the complexity of user intentions when searching the catalogue should be noted. A user who does a search by title of a known item to get the call number because s/he wishes to explore or browse the shelves around the known item, for our purposes this user is still conducting a ***known item search 1***, but the intentions of this user may be different from a user wishing to take home the physical copy of the known item.

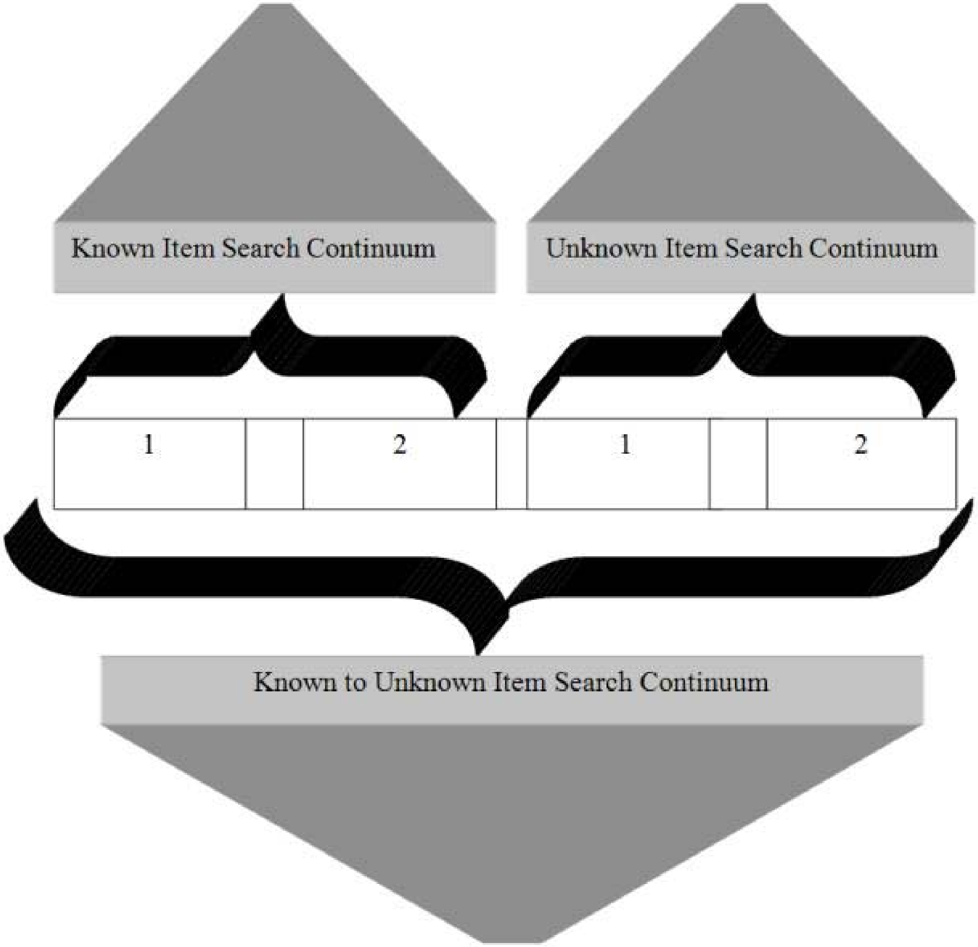
Known Item Search Continuum
Unknown Item Search Continuum
1
2
1
2
Known to Unknown Item Search Continuum

Figure 1: Known to unknown item search continuum, with two sub-continuums: the known item search 1 and 2 continuum and the unknown item search 1 and 2 continuum.

In ***known item search 1***, the user has complete and accurate information about the known item, giving this user maximum certainty going into the search of the index or catalogue. For this search, the primary purpose of the index or catalogue is to show the user whether or not the library owns the known item and, if so, where, by the classification code, the item can be physically located; either in the library where the user is at that moment (Lancaster and Joncich 1977: 19), or in some other location by inter-library loan (Svenonius 2000). By typing in all the attributes of the known item in the query, the system, in theory, produces a one-item long results list, from which the user can immediately write down the item location code. If the item is not in the database the results list will be empty. (We make the assumption that the index or cataloguing record contains no mistakes and that the retrieval system does not provide to the user approximate citations in the results list.) For this search, therefore, Cutter's finding first objective is 100% operative while the second and third objectives are, in theory, not needed. This search can be termed a *perfect match search*.

In ***known item search 2***, the user has a state of mind based on the knowledge or belief that the sought for item exists (Lee *et al.* 2006), but the user has incomplete or inaccurate information about the known item (for an example of this type of search, see Matthews *et al.* 1983: 90; see also, Dwyer *et al.* 1991; Lewis 1987; Swanson 1972). Uncertainty occurs in the user the moment the system responds to the incomplete or inaccurate information by producing a set of alternative items that it deems to be approximate to the user's input of information about the known item. The user must then identify the sought for item from this set of like-items by looking at the item's characteristics on the catalogue or index record. The assemblage '*enable[s] a user to identify a document uniquely and thus distinguish [it] from similar ones*' (Borgman 2000: 74). When the user has incomplete or inaccurate information about the known item, the index or catalogue record is particularly designed for facilitating the matching process between the user's mental image of the known item and the record of the item in the index or catalogue. The record assembles not only all the necessary identification attributes or dimensions of the item in a formulaic fashion, it often does this in a more complete and more accurate manner than in the actual physical item by bringing this information from external sources. The user quickly identifies his or her gaps and inaccuracies about the known item. This search can be termed a *best match* *search*, meaning the user must identify and select the record that best (not perfectly) matches his or her original inaccurate or incomplete mental image of the known item. ***Known item search 2***, where the user has incomplete or inaccurate information about the known item, is extremely common.

In ***unknown item search 1***, the user knows how to describe what s/he is looking for, therefore enabling this user to formulate a precise and efficient query; the traditional start state for utilizing the index or catalogue. An example of this type of search is: Who is the governor of Alaska? There is a precise answer for this type of search, the general parameters of which the user has in mind before the search. (It has even been argued that this type of subject or topic search is in fact a form of known item search. According to Bates (1998: 1186), '*knowledge specifically of what is wanted [leads to] a "known-item" search*'.) Again, as there will be more than one item in the results list, this is a best match (not perfect match) type of search.

In ***unknown item search 2*** and contrary to the previous three search-types, users only know '*fringes of a gap in [their] knowledge*', making it extremely difficult for them to identify and describe the information gap or need (Bates 1998: 1186). Because these users do not have knowledge of their information need when the search begins (Belkin *et al.* 1982; Borgman 2000), they cannot identify an effective start state from which to form a query and to use the catalogue effectively. Of all the four searches described in Figure 1, the user in ***unknown item search 2*** has a mental image of the search that is, in fact, in an '*asymmetrical*' relationship with the catalogue or index (Bates 1998: 1186). This is also a best match type of search, producing a results list that is long and especially difficult for the user to understand or process.

Concept development: *known item search 1* versus *known item search 2*

In this section, for the purpose of further defining the concepts for our developing model, we compare perfect match search, represented by ***known item search 1***, with best match search, represented by ***known item search 2***. As the only perfect match search among the four search types, ***known item search 1*** provides a theoretical baseline for the concepts that make up our index model. ***Known item search 2*** is chosen because it is the entry level best match search.

Based on the search definitions in the previous section, we make the following statements comparing the perfect match ***known item search 1*** with the best match ***known item search 2***:

- ***Known item search 1*** is, in theory, totally focused on Cutter's finding objective 1, while best match ***known item search 2*** brings into play Cutter's third selection objective.
- Therefore, the role of Cutter's second identification objective is greater for ***known item search 2*** than for ***known item search 1***.

We illustrate these two statements in Figure 2, where Cutter's identification second objective is shown as a sliding scale between Cutter's finding first objective and Cutter's selection third objective. The identifying scale exerts more power for best match ***known item search 2*** and much less power the closer the search is to the theoretical notion of a perfect match ***known item search 1***. We also associate the user's uncertainty with the identifying scale. The higher the user uncertainty about the search the greater the utilization of the identifying scale.

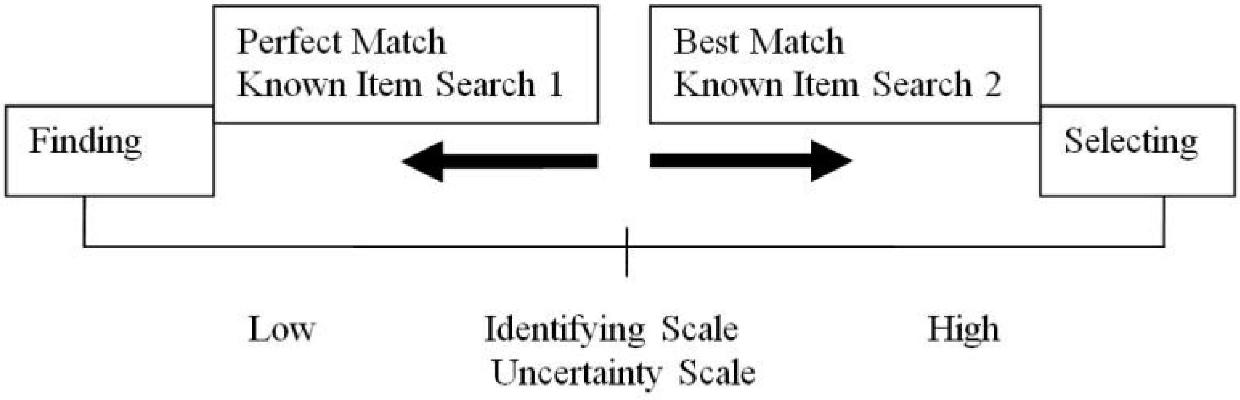

Figure 2: An illustration of the dynamic scales of Cutter's three objectives for the perfect match *known item search 1* versus the best match *known item search 2*.

In Figure 3, we translate Figure 2 into a phase or process diagram. On the left-hand side of the diagram, we represent the perfect match process of ***known item search 1***, while on the right-hand side of the diagram, we represent the best match process of ***known item search 2***. For the perfect match process, selection is immediate. For the best match process, on the other hand, selection requires intervening phases triggered by a results list with greater than one item. The user's uncertainty rises as this user realizes, from looking at the results list, that s/he has incomplete or inaccurate information about the known item being sought.

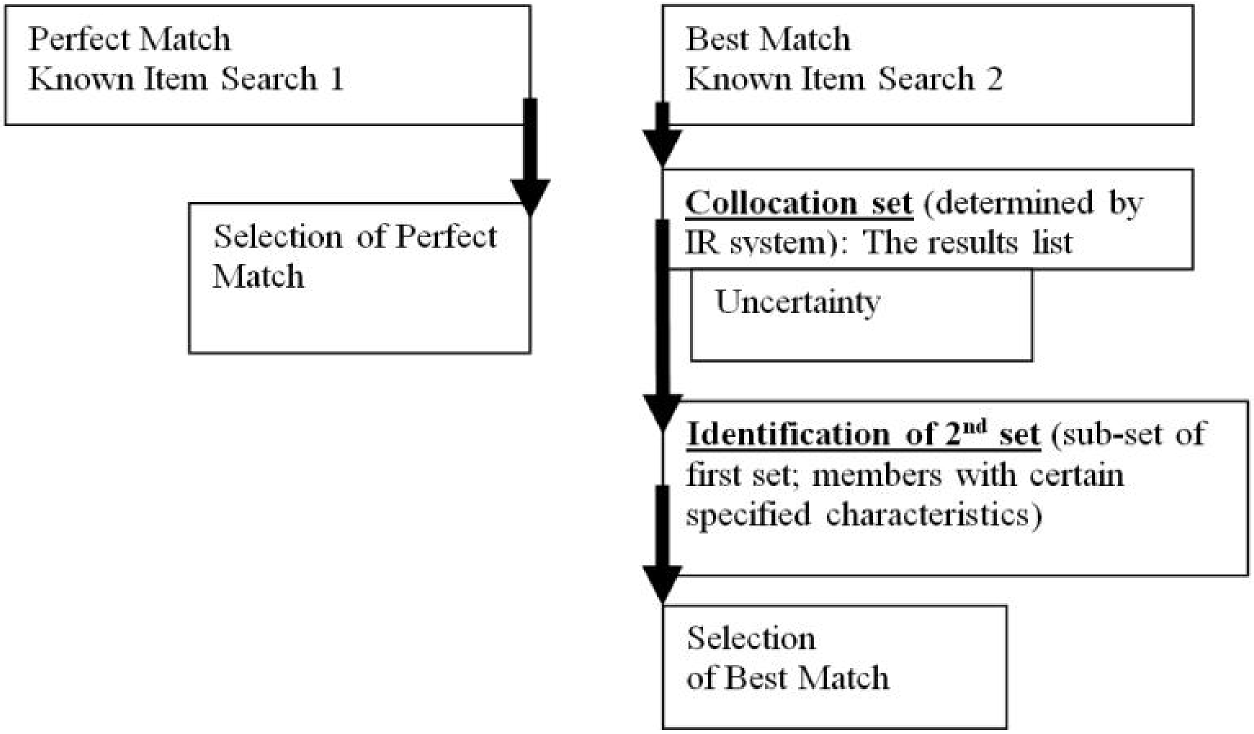
Perfect Match
Known Item Search 1
Selection of Perfect Match
Best Match
Known Item Search 2
Collocation set (determined by IR system): The results list
Uncertainty
Identification of 2nd set (sub-set of first set; members with certain specified characteristics)
Selection of Best Match

Figure 3: The difference between perfect match (*known item search 1*) and best match (*known item search 2*) in the selection process.

Figure 3 introduces the concept of the second set created by the user. In our previous expanded definition of Bush's selection concept, when '*the user first examines a* ***set*** *of items*', it is now clear that the set is the results list, which is the collocation set brought together by the information retrieval system in response to the user's query and that when the user '*identifies or recognizes*' from this set those members with '*certain specified characteristics*' that there is actually a second set created which is smaller than the first set.

Example: Oliver Twist and the second collocation set

We now further develop our model by illustrating ***known item search 2*** with the example of a user who has incomplete or inaccurate information about the item searching for the first edition of Charles Dickens' Oliver Twist. The user is surprised to see there are two items in the results list, one of which is a reference to Oliver Twist published in serial form in the journal Bentley's Miscellany in 1837. The second citation is for the first edition of the book published in 1838. With the appearance of asymmetry or noise between the user's mental image of the known item and the actual record of it in the index or catalogue, the situation shifts from a perfect match search to the imperfect world where the user will have to settle for a best match, i.e., the match that will most probably fulfill his or her information need.

Traditional, system-oriented information retrieval design assumes this user is adequately served by the results list and, based on a static conception of information need, that s/he will select the book first edition of Oliver Twist published in 1838. But for these users, for whom

Bush designed the memex machine, their associative thinking has moved on, complicating matters by forming another, second collocation set. This second collocation set is formed in the user's own mind and is not controlled by the retrieval system. Let us hypothesize that the second collocation set can be put into question form, producing the following four questions in the user's mind:

1. Are the serial and first book editions of Oliver Twist different?
2. If so, how are the two editions different?
3. Why did Dickens revise the book edition as it says in the description of the item?
4. Should I ignore my questions and just make the easiest selection, the book first edition?

These four item-questions constitute the four members of a new, second collocation set, created from the user's associative thinking. The user will either select questions 1, 2 or 3 and engage in new information seeking behaviour to find answers to the selected question, or s/he will select question number 4 and revert back to the index or catalogue-created first collocation set to obtain the location code for the book first edition.

The role of the associative index is to facilitate the user assembling the second collocation set, then identifying, recognizing and distinguishing the needed item from the other members of the set. The user can now make the selection of the new information need from the set.

The associative index versus AquaBrowser

The conceptualization of the associative index just described is fundamentally different from AquaBrowser's word cloud technology, which represents, in visual form, the concept terms the AquaBrowser system associates, by co-occurrence analysis of terms in the bibliographic citations, with the user's query. The concepts shown the user by AquaBrowser, therefore, are associated with the user's original information need.

Our associative index, on the other hand, conceives of the results list as a trigger of new associative thoughts in the user, which, according to associationism theory, are linked together in the user's memory in an associative network. These triggered thoughts constitute a second collocation set, from which the user will select a new information need. The new information need supersedes the original information need. We illustrate this fundamental difference in Figure 4.

Figure 4 is divided into two halves, labelled A and B. A, the upper half of the figure, represents our associative index conception of what is going on with the results list, while B, the lower half of the figure, represents AquaBrowser's view of what is going on with the user when s/he is reading the results list. In A, the results list triggers associative thoughts in the user's associative network; the user identifies the associative thoughts which constitutes the second set, from which the user selects a new, second information need. The second information need may be entirely different from the initiating information need in the user's query.

In B, AquaBrowser has the initiating information need, *information need 1* in Figure 4, still controlling the user's interaction with the results list. Associative thoughts of the user triggered by the results list that are not directly associated with the initiating information need are considered noise. For ***known item search 1 and 2*** and ***unknown item search 1***, indicated in the figure by K1, K2 and U1 respectively, the user's information need is strong enough to shine through the noise. However, for ***unknown item search 2***, which is an exploratory subject search, indicated in the figure by U2, the information need is so weak that the results list and its noise produce information overload.

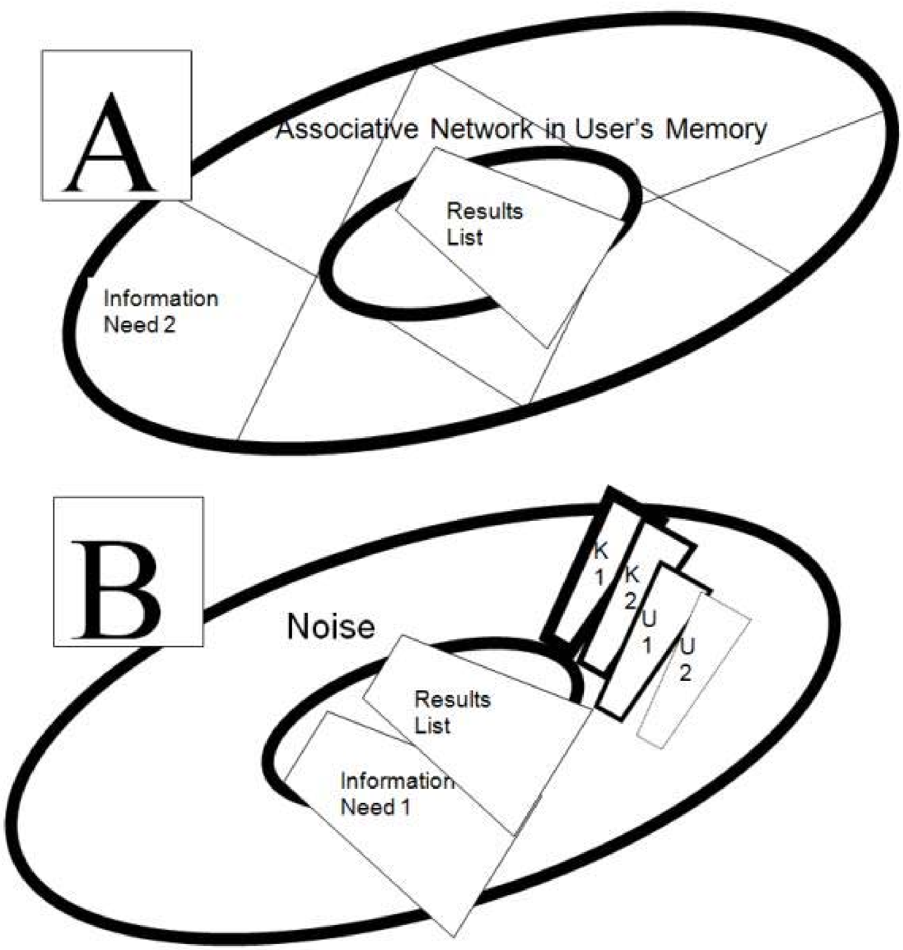

Figure 4: Two different conceptions of the user's mindset when looking at the results list: the user's mindset (circular rings), the user's information need and the results list. A. Upper half of figure is our associative index model. B. Lower half of figure is the AquaBrowser.

Associative index model for hyperlink Internet search

Remember that ***known item search 2*** is an entry level (or easiest example) for best match search, which includes the far more complicated ***unknown item search 2***. We are now ready to build the identifying, selecting and finding elements of ***known item search 2*** into a model for hyperlink Internet search, called the *associative index model*. The central feature of the model is specifying with some degree of precision the identification and selection phases of information search as they are sketched out in Bush and Cutter, specifically through the introduction of the concept of the second collocation set. The establishment of the second collocation set, which is the user's associative thinking triggered by looking at the results list, is the fundamental innovation of the model.

In our Oliver Twist example, the results list of two items constituted the information retrieval system-determined first collocation set. For our model, however, we have to transition to another type of index, forcibly one that is dependent on the user's own associative thinking, which in the Oliver Twist case consisted of the user's own thought associations triggered when the user examined the two records in the first system-created collocation set. This transition step in indexing, we believe, is at the heart of the difference between traditional cataloguing and indexing schemes and Bush's innovative, forward looking conceptualization of an associative indexing approach for the memex machine.

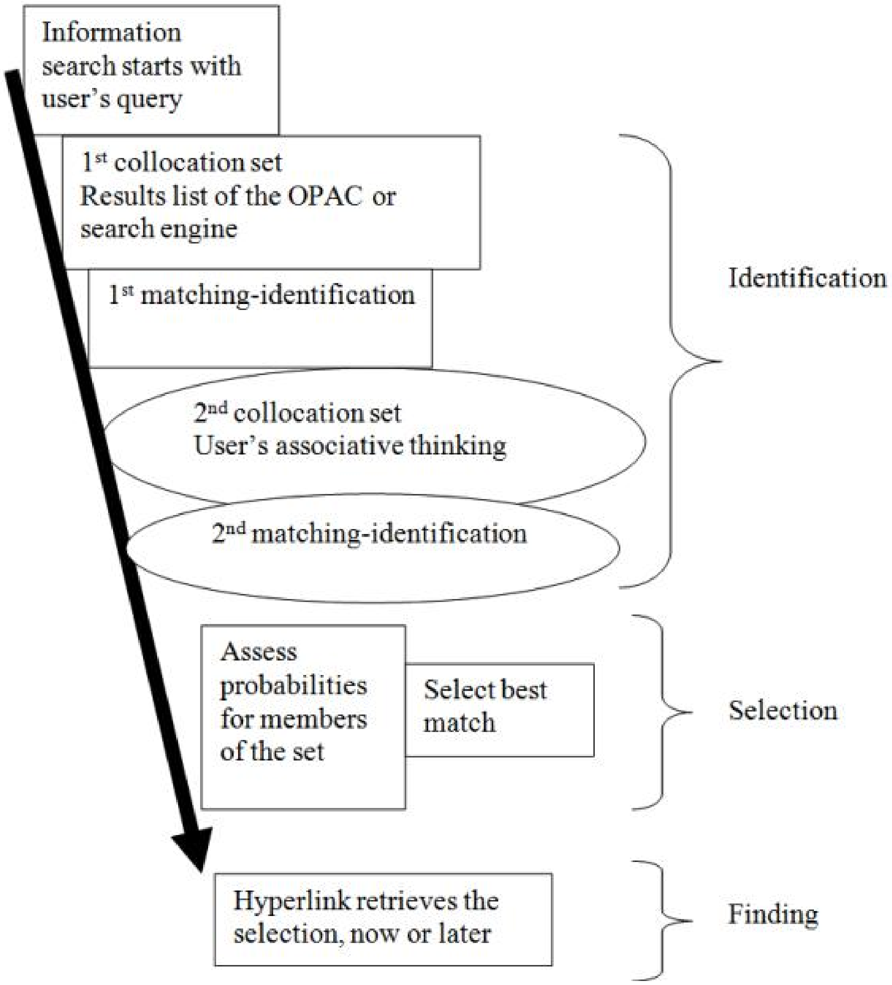
Information search starts with user's query
1st collocation set Results list of the OPAC or search engine
1st matching-identification
2nd collocation set User's associative thinking
2nd matching-identification
Identification
Assess probabilities for members of the set
Select best match
Selection
Hyperlink retrieves the selection, now or later
Finding

Figure 5: The *associative index model* for information search on the Internet.

The *associative index model* shown in Figure 5 starts with the user's initial query to the information retrieval system (e.g., the online catalogue or an Internet information search engine). The system collocates the system response set in the results list, producing the **first collocation set.**

In the **identification** phase, users examine the first collocation set, matching items in the results list to their mental image of the item being sought. In our model, the user has incomplete or inaccurate information about the sought for item, creating a results list that is more than one item long. The rise in user uncertainty produces the mental operations that follow.

The **first matching-identification**: The user must consider all items in the results list as possible contenders for the sought for item, matching a mental image of the needed item with the results list. The user utilizes certain specified characteristics to distinguish good from bad members of the set. These specified characteristics constitute the user's initiating information need, but looking at the results list also triggers associated thoughts.

The **second collocation set** is at the heart of the model. The user's consideration of the first collocation set (the results list) triggers associative thinking. These are the user's thoughts from the user's associative memory network. They may or may not be related to the user's original or initiating information need as expressed in the initial query to the system. The associated thoughts constitute the second collocation set.

The **second matching-identification**: The retrieval system asks users to give identity to the associative thoughts or associations by listing all their thoughts associated with the act of viewing the results list. The case study in the next section illustrates the listing. The identified associated thoughts constitute the specified characteristics of the user's information need.

In the **selection** phase, the user must decide which association is the most important of all the associations in the second collocation set. An association is a characteristic of the user's information need at that moment (when looking at the results list). The system can facilitate this user task by asking the user to assign probabilities to each member of the second collocation set (see the case study that follows for how this can be done). Then the highest probability member is the item that is selected. It forms the revised query to the system. In this way, as Bush originally defined associative indexing, '*one item*—[i.e., the most important associated thought triggered by the item] *selects another* [the next item]'.

In the **finding** phase, the user physically obtains the selected item by inputting a new query based on the highest probability member of the second collocation set or, if the highest probability member of this set is a hyperlink in the results list, the user clicks on the selected hyperlink to the actual item in the electronic database.

The model illustrated and tested: a case study

The case study presented here is intended to illustrate and test the functioning of the model shown in Figure 5, by comparing it with the online public access computer of the Queens University Library (New York) . This was the geographically closest catalogue equipped with the popular word association and visualization information retrieval technology called the AquaBrowser, which is used by over 700 libraries worldwide. In addition to the usual catalogue results list, the AquaBrowser shows the user word clouds, a visualization of words AquaBrowser's algorithm associates with the user's query. Users may click on the hyperlink of a concept term in the word cloud to refine their query.

Case study

On May 15, 2008, a Master's programme student in the Faculty of Education at McGill University was asked about an information need for a term paper for a course he was taking at the time of the interview. The subject was taken to the Queens University Library catalogue and asked to think about an information need he had concerning his term paper, to formulate the need into a one, two or three word query and to type the query into the online catalogue search box.

The student typed the words '*dropout rates*' into the search box, which produced a results list of seven items. We labelled this results list 'Results List 1'. On the left-hand side of the screen was AquaBrowser's visualization of a network of words its algorithms associated with the subject's query, called a word cloud. The subject was directed to select a word association from the word cloud that interested him. The subject selected the word 'cause' and clicked on it which produced a results list of 5,000 items. We labelled the first page of this results list, containing ten citations, 'Results List 2'.

Results Lists 1 and 2 were printed and shown to the subject, who was asked to look at the citations carefully from both lists. He was then asked to write down four information questions that came to mind as he looked over the citations, four questions he wanted information from the catalogue to answer. His four questions were:

1. Canadian and American dropout rate comparison: What are the similarities and differences in measurement, actual rates, socio-political determinants, projections?
2. What are the politics behind reporting dropout rates in Canada?
3. What are the most accurate measures of dropout rates in Canada and where can we find these figures?
4. What is the policy and programme significance of dropout rate research in Canada?

The subject was asked to rate the probability for each of the four questions that it would be in the final version of his term paper, or its importance to his term paper. He rated the questions as shown in Table 1.

Question	Subject's rating (in percent probability)
1	3%
2	20%

3	65%
4	12%

Table 1: Subject's rating of each of four questions in terms of probability question would be in final version of term paper.

The subject was then asked to focus on his highest rated question, Question 3 and to formulate a one, two or three word query from this question. With Results List 2 still on the screen, the subject tried and rejected several queries without pushing enter, then the subject typed the new query 'education statistics method'. He said the words in the visualization for Results List 2 had been helpful in coming up with these new search terms. The query 'education statistics method' produced thirty-seven citations. The first page of ten citations was printed out and labelled 'Results List 3'.

The three results lists produced during the interview were then put in front of the subject in shuffled order and he was asked to rank each according to a five point Likert Scale, with 1 = bad and/or not interesting and 5 = good and/or interesting. He ranked Results List 1 with a 4, 'Results List 2' with a 1 and 'Results List 3' with a 5, as shown in Table 2.

Results List (three shown to subject)	Subject's rating (on Likert Scale)
1	4
2	1
3	5

Table 2:Subject's rating of three results lists according to a 5-point Likert Scale

The researcher asked the subject why he had rated the results lists as he did. For Results List 1, the subject said it had:

> ...some [citations] related to research, but primarily American so generic search terms would produce the American context mostly. Few of the citations [from this list, however] are generic enough to use. Perhaps some could be used as a comparison study [with his Canadian information on the topic] in a small part of my paper.

For Results List 2' for the query 'cause', the subject said:

> None of these entries are relevant to research or the paper. I am surprised to see that 'cause' produced a citation list none of which had anything remotely to do with original search [query he had typed in search box].

For Results List 3, the subject said there are:

> ..several texts which I will consult [here]. They have a very specific focus on statistics, which is 'major' part of paper. Several books on statistics and methodology equal 'exactly' what my paper is concerned with. Even if they are U. S. books, they are theoretical so I can use them.

When asked to mark the citations he would later consult from Results List 3, the subject marked four of the ten citations and was given a copy of these results so that he could, if he wished, find them via their call numbers.

Discussion

The case study was designed to illustrate the model shown in Figure 5 and to indicate, in a preliminary way, the efficacy of the model in indexing the student's associative thinking while the student is engaged in matching his mental image of the sought for item with the system-produced first collocation set in the results list. As in all indexing schemes, the ultimate purpose of the *associative index* is to facilitate the student's formulation of the query to the information retrieval system, which most accurately reflects the student's real information need.

The case study illustrated how the model may work by taking the subject through the collocation/identification, selection and finding phases of the model. The model succeeded in collocating the subject's associative thinking into a second collocation set. The subject easily assigned probability values to each member of the second collocation set, then selected the highest probability question as the associative thought. This question represented the subject's most important associated thought when looking at Results Lists 1 and 2 and this thought then selected the next item, put into query form as: 'education statistics method'; the query that caused the system to produce Results List 3.

Evidence from the case study of the efficacy of the *associative index model* is anecdotal. However, the response of the subject provides some indication that it may have facilitated the student getting at his information need at that moment. Evidence for this is:

1. His oral opinion of Results List 3, which indicated that he had not seen these items and that he thought they may be useful to him in writing his term paper, whereas he had already seen the citations in Results List 1 and he did not like the citations in Results List 2.
2. This resulted in him being able to find more useful citations from Results List 3 than from Results List 1.
3. He also gave a slightly higher rating to Results List 3 than to Results List 1.

The query that produced Results List 2 was the hypertext term 'cause' which the subject clicked on from AquaBrowser's word cloud. The subject's negative reactions to Results List 2, whilst at the same time expressing a positive view of the word associations in the word cloud, indicates it is possible the case study subject had in mind a completely different interpretation of the term 'cause' than the system's matching algorithm.

Conclusion

The problem of indexing the results list in the hypertext Internet search environment is crucial to making information search easier and more natural for today's users. In this paper, the *associative index model* is an attempt to build on the principles and concepts of indexing/cataloguing in use over the last 135 years to produce a new, evolved index for searching and finding needed information the way humans naturally think.

We took as a starting point Vannevar Bush's article As We May Think, where he describes his conceptualization of an information storage and retrieval machine called the memex that he designed to give informational support to the associative thinking of the user while that user is reading a text on the memex screen. The user, according to Bush, selects the next information item s/he will need to look at based on the associated thoughts that come into the user's head, triggered by what is seen on the screen. The *associated index model* indexes the user's associative thinking at this crucial moment when s/he is about to select the next item.

Bush's memex machine is based on associationism theory. Current applications of associationism such as Google's Wonder Wheel and Aquabrowser's Word Cloud suggest system-selected associations to the user's initial query, but this, we believe, misses the point of associationism. The user's associative thinking is *ad hoc* and user specific and will not necessarily be facilitated by system-generated associations at the moment when the user is perusing the results list. This assertion, however, requires further testing.

We believe the difficulty users have with system-generated concept/word associations is not so much the trouble users have relating to the concept terms (especially domain novice users unfamiliar with a domain's conceptual and structural framework), but rather the way the human mind naturally works at a given moment when confronted with environmental stimuli. While the user may be amenable to using an index or catalogue/classification scheme before starting the search, when s/he is in a learning mode, the user is in a different mindset after the search starts and the results list appears. The results list stimulates a whole host of associative thinking in the user which, we contend, is a positive step in getting users closer to their real, underlying information need. It is not useful, except from a system point of view, to pull the user back to the original, initiating query. We have tried in this paper to disassociate these two very distinct index perspectives, particularly in the case study.

The case study illustrated and tested the *associative index model*, showing that it can be easily and practically applied in a real information search situation for a user who has a real information need. The eventual goal of our research programme is to expand the scope of our evaluation, to test the model in a defined user population in a naturalistic setting by randomly assigning members of a target population to either a control group, which receives a standard, AquaBrowser-type interface when viewing the results list, or the model group, which receives the intervention similar to the one shown in the case study. The most effective dependent variable in this case, where the information retrieval task is for a real-life student task, is the mark given to the student assignment by the course instructor. The idea of the *associative index model* will only be taken up by mainstream information search engine developers if it is shown to be effective through gold standard research design which demonstrates dramatically improved performance for real users in the real world.

Acknowledgements

The authors wish to acknowledge the financial contribution to the research of the Social Science and Humanities Research Council (SSHRC) of Canada, File No. 410-2002-1394.

About the authors

Charles Cole is a researcher, affiliated member at the School of Information Studies, McGill University, Montreal. He is also an information design consultant (Colemining Inc.). He received his Ph.D. in Information Science from the University of Sheffield, UK, his MLIS from McGill University and his B.A. in history-geography from McGill University. His current research interests include the information seeking of school children He has published extensively in JASIST, IPM and other journals. He can be contacted at charles.cole@mcgill.ca

Charles-Antoine Julien is completing his PhD in information studies from McGill University. His current main interests are links between information organization, information retrieval and information visualization. He completing an applied sciences Masters in educative technologies (Polytecnique, Montreal) and an engineering undergraduate (Polytecnique, Montreal). He can be contacted at charles.julien@mail.mcgill.ca

John E. Leide is past Associate Professor at the School of Information Studies, McGill University, Montreal. He received his PhD in Library Service from Rutgers University, his MS (Library Science) from the University of Wisconsin (Madison) and his BS (mathematics-humanities) from the Massachusetts Institute of Technology. His research interests include library and information organization, cataloguing and classification. He can be contacted at john.leide@mcgill.ca

References

- Anderson, J. R. & Bower, G. H. (1973). *Human associative memory.* Washington, DC: V. H. Winston & Sons.
- Baker, S. L. & Lancaster, F.W. (1991). *The measurement and evaluation of library services.* Arlington, VA: Information Resources Press.
- Bates, M.J. (1989). The design of browsing and berrypicking techniques for the online search interface. *Online Review*, **13** (5), 407-424.
- Bates, M. J. (1998). Indexing and access for digital libraries and the Internet: human, database and domain factors. *Journal of the American Society for Information Science,* **49**(13), 185-1205.
- Bawden, D. (1986). Information systems and the simulation of creativity. *Journal of Information Science*, **12**(5), 203-216.
- Belkin, N.J., Oddy, R.N. & Brooks, H. M. (1982). ASK for information retrieval. Part I: background and theory. *Journal of Documentation*, **38**(2), 61-71.
- Borgman, C. L. (2000). *From Gutenberg to the global information infrastructure.* Cambridge, MA: The MIT Press.
- Brooks, D. (2009, May 29). The empathy issue. *The New York Times*, p. A25. Retrieved 19 August, 2010 from http://www.nytimes.com/2009/05/29/opinion/29brooks.html (Archived by WebCite® at

http://www.webcitation.org/5s623piCt)

- Buckland, M. K. (1992). Emanuel Goldberg, electronic document retrieval and Vannevar Bush's memex. *Journal of the American Society for Information Science*, **43**(4), 284-294.
- Burke, C. (1992). A practical view of memex: the career of the rapid selector. In J. M. Nyce & P. Kahn (Eds.). *From memex to hypertext: Vannevar Bush and the mind's machine* (pp. 145-164). Boston, MA: Academic Press.
- Burke, C. B. & Buckland, M. K. (1994). *Information and secrecy: Vannevar Bush, ultra and the other memex.* Metuchen, NJ: Scarecrow Press.
- Bush, V. (1945). As we may think. *Atlantic Monthly*, **176**(1), 101-108. Retrieved 19 August, 2010 from http://www.theatlantic.com/magazine/archive/1969/12/as-we-may-think/3881/ (Archived by WebCite® at http://www.webcitation.org/5s62Fj0l6)
- Bush, V. (1991). As we may think. In J. M. Nyce & P. Kahn (Eds.). *From memex to hypertext: Vannevar Bush and the mind's machine*, (pp. 85-110). Boston, MA: Academic Press.
- Cao, J., Liang, J., & Lam, J. (2004). Exponential stability of high-order, bidirectional, associative memory, neural networks with time delays. *Physica D: Nonlinear Phenomena*, **199**(3-4), 425-436.
- Cronin, B. (2007). Introduction. *Annual Review of Information Science and Technology*, **41**, vii-x.
- Cutter, C.A. (1904). *Rules for a dictionary catalog* (4th. ed.). Washington, DC: US Government Printing Office. Retrieved 19 August, 2010 from http://digital.library.unt.edu/ark:/67531/metadc1048/m1/1/ (Archived by WebCite® at http://www.webcitation.org/5s66uVQnF)
- Dwyer, C. M., Gossen, E. A. & Martin, L. M. (1991). Known-item search failure in an OPAC. *RQ*, **31**(2), 228-236.
- Hert, C. A., Jacob, E. K. & Dawson, P. (2000). A usability assessment of online indexing structures in the networked environment. *Journal of the American Society for Information Science*, **51**(11), 971-988.
- Houston, R. D. & Harmon, G. (2007). Vannevar Bush and memex. *Annual Review of Information Science and Technology*, **41**, 55-92.
- Ingwersen, P. & Järvelin, K. (2005). *The turn: integration of information seeking and retrieval in context*. Dordrecht, The Netherlands: Springer.
- Jacob, E.K. (2004). Classification and categorization: a difference that makes a difference. *Library Trends*, **52**(3), 515-540.
- Lancaster, F.W. & Joncich, M.J. (1977). *The measurement and evaluation of library services.* Washington, DC: Information Resources Press.
- Lee, J.H., Renear, A. & Smith, L.C. (2006). Known-item search: variations on a concept. *Proceedings of the 69th Annual Meeting of the American Society for Information Science and Technology*, **43**(1), 1-17.
- Lewis, D. W. (1987). Research on the use of online catalogs and its implications for library practice. *Journal of Academic Librarianship*, **13**(3), 152-156.
- Matthews, J. R., Lawrence, G. S., Ferguson, D. K. & Council on Library Resources. (1983). *Using online catalogs*. New York, NY: Neal-Schuman.
- Meadow, C.T., Boyce, B.R., Kraft, D.H. & Barry, C. (2007). *Text information retrieval systems.* Bingley, UK: Emerald Group Publishing.
- Nelson, T.H. (1991). As we will think. In J.M. Nyce & P. Kahn (Eds.). *From memex to Hypertext. Vannevar Bush and the mind's machine* (pp. 245-260). Boston, MA: Academic Press.
- Nyce, J.M. & Kahn, P. (1989). Innovation, pragmaticism and technological continuity: Vannevar Bush's memex. *Journal of the American Society for Information Science*, **40**(3), 214-220.
- Otlet, P. (1934/1989). *Traité de documentation*. [Treatise on documentation.] Liège, Belgium: Centre de lecture publique de la communauté française.
- *Oxford dictionary of current English*. (1985). Oxford: Oxford University Press.
- Pirolli, P. (2007). *Information foraging theory: adaptive interaction with information*. Oxford: Oxford University Press.
- Plotkin, H. (2004). *Evolutionary thought in psychology: a brief history*. Malden, MA: Blackwell.
- Popper, K. (1975). *Objective knowledge: an evolutionary approach.* Oxford: Clarendon Press.
- Raaijmakers, J.G.W. & Shiffrin, R.M. (1981). Search of associative memory. *Psychological Review*, **88**(2), 93-134.
- Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, **85**(2), 59-108.
- Smith, L.C. (1991). memex as an image of potentiality revisited. In J. M. Nyce & P. Kahn (Eds.). *From memex to hypertext: Vannevar Bush and the mind's machine*. (pp. 261-286). Boston, MA: Academic Press.
- Svenonius, E. (2000). *The intellectual foundation of information organization*. Cambridge, MA: The MIT Press.
- Swanson, D.R. (1972). Requirements study for future catalogs. *Library Quarterly*, **42**(3), 302-315.

- Tebbutt, J. (1999). User evaluation of automatically generated semantic hypertext links in a heavily used procedural manual. *Information Processing & Management*, **35**(1), 1-18.
- University of California Libraries. *Bibliographic Services Task Force.* (2005). *Rethinking how we provide bibliographic services for the University of California*. Oakland, CA: University of California. Retrieved 26 March, 2010 from http://libraries.universityofcalifornia.edu/sopag/BSTF/Final.pdf
- Wildemuth, B.M. & O'Neill, A.L. (1995). The 'known' in known-item searches: empirical support for user-centered design. *College & Research Libraries*, **56**(3), 265-281.

How to cite this paper

Cole, C., Julien, C.A. & Leide, J.E. (2010). "An associative index model for the results list based on Vannevar Bush's selection concept" *Information Research*, 15(3), paper 435. [Available from 21 August, 2010 at http://InformationR.net/ir/15-1/paper435.html]

Find other papers on this subject

Scholar Search | Google Search | Bing

Bookmark This Page

Contents | Author index | Subject index | Search | Home