Identifying Clusters of User Behaviour in Intranet Search Engine Log Files

Dick Stenmark
IT University of Gothenburg, Department of Applied IT, S-41296 Gothenburg, Sweden
E-mail: dick.stenmark@ituniv.se

When studying how ordinary Web users interact with Web search engines, researchers tend to either treat the users as a homogeneous group or by grouping them according to search experience. Neither approach is sufficient, we argue, to capture the variety in behaviour that is known to exist amongst searchers. By applying automatic clustering technique based on self-organising maps to search engine log files from a corporate intranet, we show that users can be usefully separated into distinguishable segments based on their actual search behaviour. Based on these segments, future tools for information seeking and retrieval can be targeted to specific segments rather than just made to fit the "the average user". The exact number of clusters, and to some extent their characterristics, can be expected to vary between intranets, but our results indicate that some more generic groups may exist. In our study, a large group of users appeared to be "fact seekers" who would benefit from higher precision, a smaller group of users were more holistically oriented and would likely benefit from higher recall, whereas a third category of users seemed to constitute the knowledgeable users. All these three groups may raise different design implications for search tool developers.

Introduction

In this article, we discuss whether users of a Web-based information seeking tool should be understood and analysed as individuals with unique requirements and preferences or seen as contributors to a collective behaviour that may be described using mean values and averages. We argue, although both extremes have their merits, that too often has the user been bundled with thousands of others at the expense of finer details and deeper understanding. At the same time, the analysis of thousands of individuals would be extremely resource consuming whilst results based on the examination of a handful could easily be biased. We therefore suggest a middle way, where Web search engine users' similarities in seeking behaviour are used to form clusters of users that thereafter can be analysed in depth.

A decent amount of research on how ordinary Web users interact with public search engines such as AltaVista (Silverstein, Henzinger, Marais & Moricz, 1998), EXCITE (Jansen, Spink, Bateman & Saracevic, 1998) or Alltheweb (Jansen & Spink, 2003), has been carried out over the last decade. Automatically generated log files from these systems have been studied and have generated useful statistics on the amount of time typically spent with the search tools, the average query length, the mean number of result pages requested, the use of advanced features and Boolean operators (or the lack thereof), and these studies have allowed us to notice emerging trends in user behaviour. We hence begin to know a few things about the average search engine user. However, as Cooper (1999) argued, there is no such thing as a typical user. It must be assumed that people who search for information have different levels of experience and education, diversified and personalised information needs and thus behave very differently. To only look at the average numbers would mask the diversity and richness that exist in search behaviour, we argue. Another common approach is thus to divide users in *a priori* defined groups, most notably in experts vs. novices (Moore, Erdelez & He, 2007). This is again is problematic, since these concepts are far from well-defined and based on the researchers' assumptions that there are both experienced and novice users out there and that the level of search experience should affect Web search behaviour. We instead suggest that one should look more openly at the users' real behaviour and use clustering techniques to identify and analyse the groups that naturally emerges out of such an activity, as previously done by Chen & Cooper (2001). Doing so avoids the average user syndrome and also allows us to study behaviour without being biased by expectations or assumptions.

The general understanding of a cluster seems to be that it is a group of objects whose members are more similar to each other than to the members of any other group, and clustering is thus the process of organising object into groups based on some sort of similarity between the objects so that that intra-cluster similarity is

Journal of the American Society for Information Science and Technology, 59(14), pp. 2232-2243, 2008.

high and inter-cluster similarity is low(er) (Maarek & Ben-Shaul, 1996). The focus of this paper is not on clustering *per se*; our aim is not to invent new algorithms or to advocate one algorithm over another. Instead, we contribute to the understanding of intranet search engine usage by showing that clustering of users based on behaviour is both feasible and informative. Instead of investigating common variables such as number of query terms or search session duration *one by one*, this study draws on Chen and Cooper's (2001) study and examines a large number of commonly studied Web search variables simultaneously. These variables are used to form an 11-dimensional vector for each user and Self-Organising Map (SOM) technique is used to reduce the data and project it onto a two-dimensional grid, which makes it possible to visualise the result. We thereafter cluster the data to identify segments of similar usage and qualitatively analyse the characteristics of these clusters. By identifying similarities within and differences between clusters of intranet users, we provide valuable knowledge for design of future search tools which can result in improved system performance and enhanced search quality.

We first review some of the previous work in this area (section 2) before describing our research setup, which contains a brief explanation of self-organising maps (section 3). Thereafter, we present the results of our clustering (section 4) and discuss the qualitative analysis of our findings (section 5).

Previous Work on User Conceptualisation

Marketing people have since long recognised the fact that not all people behave the same. To be able to diversify product design, marketing strategies and other efforts one approach has been to divide customers into homogeneous segments of buyers (Kotler, Armstrong, Cunningham & Warren, 1996). One of the most widely applied techniques of segmenting customers is to use various statistical and data mining methods, in particular *basket analysis*, which can be described as to determine correlations between different products placed in the same shopping basket (Berry & Linoff, 1997). In their study of recommender systems in the apparel domain, Ghani and Fano (2002) argue that such a data mining approach is relevant not only to their context but to a wider class of products and that abstracting from the product layer to attributes such as personal tastes can add a potentially valuable dimension to such systems.

The aforementioned approach has to some extent also been used in library and information science (LIS) studies. Eason, Richardson and Yu (2000) use k-means cluster analysis to identify seven distinct groups of e-journal users. Their study shows that log file analysis and clustering can successfully be combined to reveal otherwise hidden usage patterns in a both unobtrusive and realistic way. Still, not many researchers have used this approach. Chen and More note that many of the previous studies of patterns of user behaviour have sorted the users into various groups typically based on who they are (e.g., adult/ child, male/ female) or what they have (e.g., level of education/ training/ experience), and not on what they actually do. In particular, *search experience* is an attribute that has been used to differentiate between different types of users (Moore, Erdelez & He, 2007). The typical approach has been to contrast *novice* users to *experienced* users, but although numerous studies have been carried out over the years (Moore *et al.* found more than one hundred papers when doing their literature review) the results are inconsistent and sometimes contradictory. The authors identified several reasons for this problem, but much of this can be attributed to the lack of consensus on the definition of the concept experience. The authors found 19 unique concepts that were used to describe the search experience variable. It can be questioned whether experience is a useful concept when it comes to grouping users of search technology for analytic purposes.

On a broader note, Chen and Cooper (2001) point out that while the kind of subjective *a priori* grouping discussed above is effective, it is far from exhaustive. When groups are decided *a priori*, they are formed based on the researchers' subjective and possibly biased understanding of what will be useful from an analytic point of view. These assumptions may be wrong or only partly right. Instead of grouping users based on what they are or what they have, Chen & Cooper argue that we should look more open-mindedly on what they actually *do*, and let these actions form the patterns.

One way to operationalise such an approach would be to use clustering techniques along the lines used by Eason, Richardson & Yu (2000). However, since Eason and colleagues studied the use of e-journals and we want to characterise Web search engine users, we shall instead use Chen and Cooper's (2001) paper as a point of reference. Even though Chen and Cooper did not study search engines, their users were searching for information and are thus more similar to the users we study in our work.

Chen & Cooper (2001) developed a set of 47 variables pertaining to library catalogue usage. The 47 variables were reduced to 16 principal components that defined a user's Web search behaviour. Combining hierarchical and non-hierarchical clustering analysis methods, Chen & Cooper identified 6 naturally emerging clusters of user behaviour. The largest cluster (with 37% of the sessions) was labelled "Unsophisticated usage". The second largest cluster (27%) was called "known-item searching. Third, with close to 14% came a cluster called "Highly interactive with good search results". The fourth largest cluster (with 11 %) was labelled "Relatively unsuccessful usage". In fifth place came the second smallest cluster (8%) which was named "knowledgeable and sophisticated usage". The smallest cluster, finally, had only 3% of the sessions and was named "Help-intensive searching". These groups were found to be present in the same proportions also in a second sample, thus indication that the structures were not purely accidental (Chen & Cooper, 2001).

We intend to build on and extend the results reported by Chen & Cooper (2001). Our work differs from the above in two important ways. Firstly, Chen & Cooper

acknowledge that although their data came from a Web-based catalogue, the reported usage was more indicative of patterns of usage of library catalogue searching than of search engine usage. We, in contrast, have analysed log files from a Web search engine and its usage. Secondly, we have examined the behaviour of *intranet* users – a group thus far often neglected in LIS studies. Still, we intend to compare our results to that of Chen & Cooper and discuss both similarities and differences.

Research Context and Method

In the following section, we account for the search engine and the context in which it operates, and for the research approach used for this work.

The TransMech intranet

This study is based on real data from real users with real information needs. The log file was obtained from the TransMech intranet in 2004. TransMech is a European hardware manufacturer with offices and factories in many countries around the world. In 2004, there were approximately 70,000 employees in the company group, which consisted of nearly a dozen individual companies. The TransMech intranet was started in 1995 and did in 2002 consist of more than 1,500 Web servers. The exact amount of documents (or Web pages) available on the intranet was impossible to determine, but corporate officials estimated it to be in the region of 8-900,000 documents. Content was typically work-related and provided by a relatively small group of informants in a top-down fashion.

Since 1998, TransMech uses Ultraseek as their intranet search engine. Ultraseek is a commercially available keyword-based search engine that allows the use of + (plus) and – (minus) to indicate that a term MUST or MUST NOT appear in the document (instead of Boolean operators such as AND or NOT). Quotation marks are used to indicate a string search and all these features may be combined. For example, the query apple –mac "fruit salad" would mean a search for documents containing the word apple, but not the word mac and the phrase "fruit salad". Results are returned in chunks of 10 where the user may access the next chunk by clicking the "next" button.

Research method

The raw data was collected between October 14th and October 21st 2004 from TransMech's search engine as a transaction log in the combined log format. The log details include information such as IP-address, time stamp of access, and what kind of request that was made. The request part of the log entry consists of a different number of Ultraseek parameters most of which are neglected in this analysis.

The log file contained 61,679 entries. We sorted the log file on IP-address and datetime, and the number of activities from each unique IP-address was counted. The most active addresses were examined manually to identify and remove obvious proxies (i.e., servers relaying queries from multiple users). After this modification, which removed a total of 109 IP-addresses, the cleaned set contained 7,902 IP-addresses, which now were considered to represent individual users.

Even though transaction log analysis (TLA) is a well-established method (see Jansen, 2006), it must be acknowledged that no standardised metrics have been agreed upon and interpretations and definitions differ between studies (Li, Cao, Xu, Hu, Li & Meyerzon, 2005). To construct our vectors, we have collected the parameters most frequently used in TLA-based studies conducted on both the Internet and on intranets. These are presented in the next subsection.

Research Parameters and Metrics

In table 1, we offer our interpretation of the parameters identified in TLA studies of Web search engine usage.

Data Pre-processing

As Desmet (2001) points out, as the number of products (or features or, as in our case, variables) grows, the size of the distance matrix (i.e. product × product) can become very large, and this means that manual processing becomes very difficult if not impossible. Here is where automatic clustering comes in handy since it scales to the full capacity of the computer resources available. Desmet suggests and demonstrates the usefulness of Self-Organising Maps (SOM) to cluster and visualise the data through automatic processing. A SOM is thus particularly useful when data is numerous and when the distribution of the variables is unknown, and since this is exactly the case for search engine log file data, we apply the same method in this paper.

The data pre-processing and the tuning of the SOM software was carried out as part of a Master Degree project by a student under the author's supervision (see Strindberg, 2006). Using the above variables as input, an 11-dimensional vector for each of the 7,902 logged users was formed. Following previous approaches (e.g. Vesanto, Himberg, Alhoniemi & Parhankangas, 1999; Desmet, 2001), these vectors were thereafter fed into the MatLab software package. MatLab can be described as a numerical computing environment with its own programming language. The software provides easy matrix manipulation, plotting of functions and data, implementation of algorithms, creation of user interfaces, and interfacing with programs in other languages. To do the actual processing and to visualise the results, the SOM Toolbox was used (CIS, 2008).

Before the vector values could be compared and visualised their values had to be normalised. The SOM toolbox uses Euclidean metric to measure distance between vectors and without normalisation, variables with higher absolute values would have greater impact on the distances measured than would variables with lower values (Vesanto, Himberg, Alhoniemi & Parhankangas, 1999). The technique used here is the default setting of the SOM toolbox, simply scaling all vectors elements to

TABLE 1. Information-seeking variables used in this study.

1. Query length ("mean no. of terms per query"). A term is defined by Jansen et al. as: "…any unbroken string of characters (i.e. a series of characters with no space between any of the characters" (Jansen et al., 1998, p. 211). Terms thus included words, acronyms, numbers, symbols, URLs, or any combination thereof. We have followed this definition but we have chosen not to include zero length queries.
2. Number of find similar ("mean no. of clicks per session"). This is an Ultraseek feature similar to Google's Similar pages. For each document link on the result page the user can click the Find similar link to retrieve more pages similar to the particular result document. This parameter holds the number of such requests made by the user.
3. Time examining documents ("average time in seconds"). In Jansen and Spink's (2003) study, they calculated the time each user spent reading a retrieved document as the time from clicking on the link until returning to the search engine. We copy that definition.
4. Time examining result pages ("average time in seconds"). When the user is presented the result page there are typically five actions to choose amongst. The user can identify a promising link and click on it, which would take the user away from the search engine and display the actual document. Alternatively, the user may move to the next bunch of (ten) result links by clicking on a next button. A third option would be to revise the query and resubmit. The fourth action would be to click on the Find similar link (see item 2). A fifth option would be to give up and leave the search engine. Unlike the first four, the last action does not generate a log entry. For this parameter we calculated the time from the result page was displayed to any of the first four actions was carried out.
5. Session duration ("average time in seconds"). There are at least two different ways to define session duration. The simplest one is to define the session duration as "measured from the time the user submits the first query until the user departs the search engine for the last time (i.e., does not return)" (Spink & Jansen, 2004, p. 44). This is a straightforward approach but it has been criticised as being too naïve, especially when the log file covers different days (He & Göker, 2000). A commonly used alternative is to look at the idle interval between two consecutive activities from the same user and if this interval is "long enough" consider it a session break. We have used the latter approach, often referred to as the "timeout" method (Huang, Peng, An & Schuurmans, 2004). To determine what is "long enough", we have used the approach described in Stenmark (2005) and set the idle threshold to 13 minutes.
6. Number of queries ("mean no. of queries per session"). A query is the search string entered by the user and defined by Jansen et al. (1998) as "one or more search terms, and possible includes logical operators and modifiers…" (p. 211). We have used the same definition and counted the number of queries submitted during a session.
7. Number of viewed hits ("mean no. of viewed documents per session"). If a user clicks on any of the links on a result page, they have viewed a document according to Jansen and Spink's (2003) definition. We have copied this approach.
8. Requested result pages ("mean no. of result pages requested per session"). When a user submits a query to the search engine, it typically returns a result page containing links to the (ten) best matching documents. This means that every user gets to view at least one result page. In this study, we do not count this first result page view (as it is trivial), but only explicit requests for result pages.
9. Number of activities ("mean no. of activities per session"). The number of activities is the sum of all the interactions a user can have with the search engine, plus the inclusion of the user's first view of the interface, prior to submitting the first query.
10. Number of sessions ("mean no. of sessions per active day"). This parameter holds the mean number of sessions that the user engages in during an "active day", i.e. a day when the search engine is used. Studies where the session length is not based on timeout method can, by definition, never report more than one session per day.
11. Number of active days ("no. of days"). This parameter simply shows how many days the users visited the search engine during the measured seven days. It can thus range from 1 to 7. Many studies of the public web only have data from a single day (or part thereof).

have the variance equal to 1. The next section describes our approach in general terms. Those interested in the mathematical and technical details are referred to the technical report (Strindberg, 2006) or the SOM Toolbox online manual (CIS, 2008).

Cluster generation

Our objective is not to study clustering methods *per se* but to see what new knowledge about users' search behaviour can be gained by applying clustering methods search log data. We used the k-means algorithm out of the SOM toolbox but other approaches could also have been used (see Xu & Wunsch (2005) for a useful review of different clustering methods).

The concept of self-organising maps can be described as a set of neurons organised as a fixed net of predetermined size. Each neuron is a d-dimensional weight vector where d is the dimension of the input vectors. On the output layer, the neurons are connected to their neighbours so that similar neurons will be closer together than more dissimilar neurons (Vesanto et al., 1999; Desmet, 2001). Similarity is based on the Euclidean distance as described by Vesanto et al. (1999). Many different forms of output can be generated, e.g., sheet, torus or cylinder, but typically a low-dimensional grid is

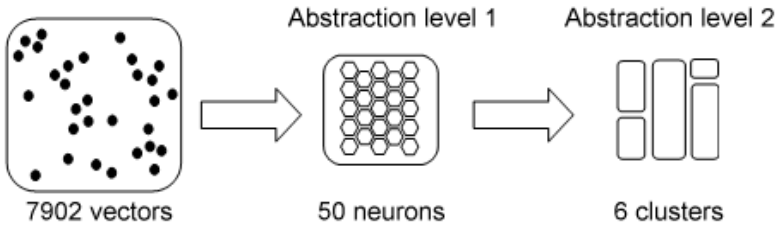

FIG 1. Going from data vectors to clusters in a two-step approach (illustration adopted from Vesanto & Alhoniemi, 2000).

chosen. For easy visualisation, we have chosen a two-dimensional sheet.

The approach used in this work is to cluster SOM rather than cluster the raw data itself, as suggested by Vesanto and Alhoniemi (2000). The primary benefits with this approach are that it significantly reduces the computational load and is less noise-sensitive. We have thus followed the two-layer approach depicted in Figure 1 below. First, the 7902 vectors were reduced to 50 neurons using the SOM algorithm (abstraction level 1). Thereafter, these neurons were arranged into 6 clusters using the k-means algorithm. The SOM-generated neurons thus served as an intermediate step. Several different clusters can be generated at abstraction level 2 and to select the "best" one, we used the Davies-Bouldin index to calculate a validity score (Davies & Bouldin, 1979). See the appendix for details.

Results

In the following section we present the results received from the SOM processing of creating six clusters by first presenting the size of the clusters and thereafter account for some of the major characteristics of each cluster.

Cluster sizes

The largest cluster in terms of number of users was cluster D (seen in the centre of Figure 2), with 32% of the users. The second largest cluster (29%) was cluster E, which is found in the upper right side of the map in Figure 2. These two clusters account for well over half of the user population. The two smallest clusters (Cluster C (5%) and Cluster F (7%)) are located in the lower left and lower right sides of the map, respectively. Being far apart in the graph means that their users have behaved very differently. Cluster A (18%) and Cluster B (9%), which together account for a quarter of the users, are co-located on the left side of the map in Figure 2.

Cluster contents

In Figure 3, we show how the values of each variable vary between clusters. The X axis shows the 11 variables: 1) Query length, 2) Number of find similar, 3) Time examining document, 4) Time examining result page, 5) Session duration, 6) Number of queries, 7) Number of viewed hits, 8) Number of requested result pages, 9) Number of activities, 10) Number of sessions, and 11) Number of active days.

The Y axis shows the spread from a normalised average and should be read variable by variable. One should, for instance, thus *not* compare variable #1 and variable #11 and believe that these two variables have similar values. Instead, one should use the figure to observe that cluster F deviates from the rest when it comes to variable #4. The six lines represent the different clusters.

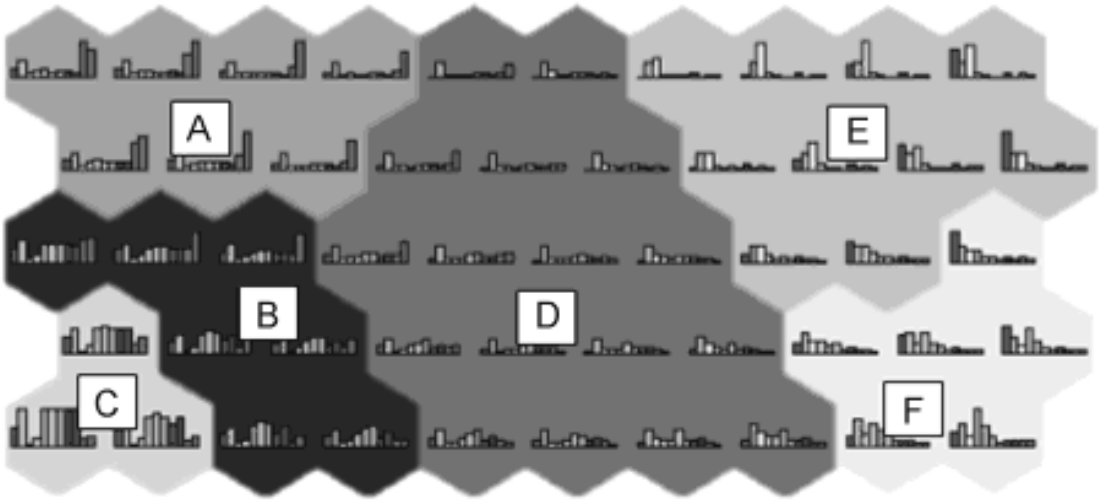

FIG. 2. Cluster map (abstraction level 2) with bar chart representing the ingoing vector components.

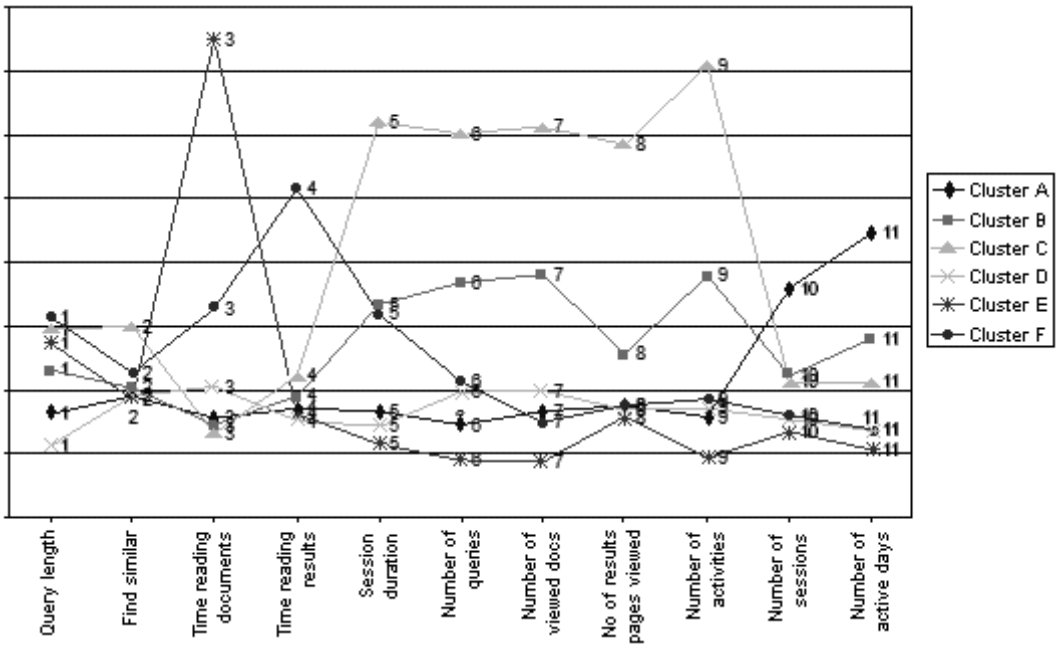

FIG. 3. Fluctuation in variable values between clusters A to F.

A number of interesting observations can be made from Figure 3. Cluster A users are characterised by their frequent use; both in number of days (variable #11) and in number of sessions per day (variable #10). These users are active three days per week and return to the search engine many times during an ordinary day for short in-and-out type of queries.

Cluster B users are not extreme in any aspect but still have a very specific behaviour. They are few, have quite long sessions (#5), ask more questions (#6) and view more documents (#7) and result pages (#8) per session than do most users, cluster C users excluded. They typically visit the search engine two days per week.

Cluster C users are characterised by the large number of activities they engage in when using the search engine one or two days per week. Their search sessions are long (#5) and they submit many queries (#6), browse through many documents (#7), request many result pages (#8), and make relatively frequent use of the Find similar feature (#2). However, they spend extremely little time reading each document (#3).

Cluster D is the largest cluster and is characterised by the fact that its users have no distinguishable characteristics. They are quite similar to cluster 1 users in that they submit short in-and-out queries and look at very few results, but they do not use the tool nearly as frequently.

Cluster E is the second largest clusters and it holds the users who are the least active (variables #10 and #11). Users in this segment are similar to clusters A and D users, but use the search engine only once a week and engage in very few activities when there. What distinguishes them from the other two is that they only click on one document but spend much time reading this; twice as much as any other user category (#3).

Cluster F, finally, is one of the smallest and its users are characterised by the amount of time spent browsing through the output (#4). These users also formulate the longest queries (#1), and spend a useful amount of time reading the two or so documents they eventually chose to view (#3).

Discussion

We have argued that there is no typical intranet searcher that behaves in one typical way and it is therefore problematic to have merely one search tool with one single interface. It seems unlikely that such a setup would allow for an optimal search experience. However, it would also be impossible to let every employee have his or her own tool or his or her own interface. Is there a middle ground somewhere? We think that clusters of similar users may provide a feasible trade-off. In this study, six such clusters have emerged and been analysed. Below, we shall discuss the characteristics of these clusters, compare them to those of Chen & Cooper (2001) and look at some possible implications for design of IR systems.

Differences in Variables

First, let us look at the variables in Figure 3. Many of the lines are gathered near the bottom of the graph, which means that intranet searchers at TransMech overall have a low level of activity. Variable 1 and variable 2 are grouped rather closely together, indicating relatively homogeneous behaviour for these parameters. This means that, regardless of cluster, users submit short queries (variable 1) and do not use the find similar feature (variable 2). The average number of terms used in a query was 1.45, to be compared to the approximate 2.5 terms

per query reported for the public Web (Spink & Jansen, 2004). Chen & Cooper (2001) do not report this value.

Other variables are not as uniform. Most striking is variable 3 (time spent reading a document), which shows a huge spread where users in cluster E (and to some extent in cluster F) really stand out. Also other variables show differences depending on cluster, e.g. variable 9 (no. of activities) and variable 11 (no. of active days).

Cluster Content

Cluster A users are characterised by their frequent use. They interact with the search engine on a nearly daily basis and submit short queries. We suggest that these users are engaged in "fact-based" tasks or that they use the search engine as a navigation aid to quickly find what they look for. Bilal (2000) defines a fact-based task as one that is usually uncomplicated, requires a single, straightforward answer, and does not require research to find the answer. We call this group "fact seekers".

Cluster B users are not extreme in any aspect but still have a very specific behaviour. They are few, have quite long sessions, ask more questions and view more result pages and documents per session than do most users, cluster C users excluded. They typically visit the search engine two days per week. We suggest these be labelled "Interactive users"

Cluster C users are not active often but when they are, they use the search engine heavily. This user group is small and can be described as more extreme versions of cluster B users. This behaviour, in combination with the short document browsing duration, suggests to us that these users are engaging in information seeking not to retrieve an "answer" but to broaden their understanding of a topic. The very process of seeking may provide the learning required to satisfy the information need. We call these "Intensive searchers".

Cluster D is the largest cluster and these users are quite similar to cluster A users in that they submit short in-and-out queries and look at very few results. Cluster D users use the search engine less often and less actively. One interpretation of this is that clusters A and D users have similar training (or lack thereof) but do different tasks. We call these users "Unsophisticated users".

There are two things that distinguish cluster E users. One is that they are the least active users; the have the lowest values for variables 5-11. The other thing is that they do not bother to look at the retrieved results very long before selecting one document which they read very carefully. This could mean that they immediately identify the right document in the result set and that this document actually contains what they are looking for, perhaps due to very well-defined information needs. Still, it is probably not a simple fact they are after, but something that requires them to read the entire document. We label this cluster "Occasional users".

Cluster F users, who spend more time than any other group examining the result pages and a fair amount of time reading documents, may be more experienced searcher. The fact that they use more query terms support this hypothesis. We also suggest that they are collecting information rather than simple facts for some specific task. These are the "knowledgeable users".

Comparative Analysis

It should be acknowledged that Chen & Cooper (2001) use features from online catalogue usage rather than Web search engine usage and that their variables in many ways differ significantly from those used in this study. Consequently, one should not draw too much on the similarities and differences found between the two studies. However, it is interesting to note that both studies ended up with 6 clearly identifiable clusters of usage patterns. This may be incidental but it may also indicate that there are approximately a handful of different seeking behaviours and that these similarities go beyond individual tools or studies. Further research is needed to test this more systematically.

It is also quite telling that the largest cluster in each study could be labelled "unsophisticated users" and held approximately one third of all sessions. Users in these categories viewed few documents and did not spend much time on the results. Again, this may indicate a general pattern – many casual searchers are fairly unaware of or do not care about the more sophisticated features available in today's search tools.

Interesting is also the existence of a small group of user whose behaviour can be characterised as "knowledgeable usage". Again, these two groups are very similar in both size (8% and 7&, respectively) and content (read result pages and documents carefully) and may indicate the existence of a universal type of searcher.

The remaining four clusters were close in size but rather different in content (see table 2). For example, the help feature was used intensely by one of the online catalogue categories. The usage of the help feature was not included in the search engine study since this feature is not generally included in TLA studies.

Design Implications

Clusters A, D and E together represent a large bunch of "casual users". Fact seekers (cluster A) come often but do very little whereas unsophisticated users (cluster D) do very little and quite seldom. Both these groups appear to be retrieving facts, i.e., they are looking for a single, straightforward answer. Based on studies of children's use of search engines for fact-based tasks, Bilal (2000) suggests that result link descriptions can assist searchers in making better navigational decisions. It can be argued that all searchers would benefit from relevant descriptions, and we concur, but it also seems plausible that fact seekers who do not appear to read very carefully would benefit more than other user categories.

In addition, fact seekers such as users in groups A and D are also likely to benefit from high precision (i.e., the percentage of retrieved documents that are relevant) but not necessarily from high recall (i.e., the percentage of relevant document in the entire collection that were actually retrieved).

TABLE 2. Comparing the clusters found in Chen & Cooper's (2001) study with our clusters.

Chen & Cooper's (2001) study		This study	
Cluster label (size)	Cluster content	Cluster label (size)	Cluster content
Unsophisticated usage (37%)	Few documents viewed Relatively short viewing times	Unsophisticated usage (32%) (Cluster D)	Short in-and-out queries Look at very few results Use the search engine less often and less actively
Known-item search (27%)	Clear information need Relatively simple search	Occasional users (29%) (Cluster E)	Very few activities Quickly selects a document Reads the one document very carefully
Highly interactive usage (14%)	Longest sessions Requesting most pages Submitting most queries	Fact seekers (18%) (Cluster A)	Quick in-and-out queries
Relatively unsuccessful (11%)	Least active Shortest sessions Shortest viewing	Interactive users (9%) (Cluster B)	Quite long sessions Ask more questions and view more result pages and documents per session than do most users
Knowledge-able usage (8%)	Longest viewing per doc Most time spent reviewing	Knowledge-able usage (7%) (Cluster F)	Spend much time examining the result pages and a fair amount of time reading documents
Help-intensive searching (3%)	Heavy use of Help Fewest sessions Spent time reading each doc	Intensive searchers (5%) (Cluster C)	Not active often but when they are, they use the search engine heavily Short document browsing duration

Cluster E users are also infrequent users but not necessarily fact seekers since they read their (one) document very carefully. However, they do not seem to care to wade through many result pages so these users would probably also prefer precision to recall.

Together, clusters A, D and E represent 80% of the search engine users at TransMech. There is always a trade-off between precision and recall. If one is optimised, the other one typically suffers (Buckland & Gey, 1994). Traditionally, information retrieval systems have been designed to do well on both these measures. Here we see that a majority of the users appear to be less concerned with recall, an observation that echoes previous suggestions on Web searching (Nielsen 1999). The design implication to be derived from this finding is that for many intranet users, precision in search tools can be prioritised at the expense of recall.

Although users from clusters A, D and E are infrequent users who do not exploit the features of the search tools to their full potential, it might still be a good idea to involve members from these user categories when implementing an intranet search engine. These users will probably not expect fancy functions or ask for advanced features (since they do not seem to use them), but involving these users early in the process might help organisations understand the reasons for their low activity levels. In addition, by gaining these users' acceptance early in the implementation phase, they may be encouraged to become more effective seekers.

Also clusters B and C users show similarities; they return several times weekly and engage in quite a lot of activities. However, cluster C users are more active than are cluster B users. Both these user groups appear to favour recall over precision since it seems they are interested in a holistic view rather than an atomic answer. The challenge that high recall presents is the way the results is to be presented to the user. It is difficult to get a holistic overall picture from a list of the top 10 or so returned links, which search engines typically produce.

Bilal (2001) suggests that an interface that displays the hierarchical structure of concepts in some sort of navigational maps would facilitate users' browsing, since it allows the users to select the appropriate concepts by using recognition rather than cognitive skills. Self-organising map, such as the ones used in this study, have actually been applied to support user browsing of the taxonomies of Yahoo (Chen, Houston, Sewell, & Schatz, 1998) and may provide a way forward for interactive users (category B) and intensive users (category C).

Since users in categories B and C are active searchers, it can be expected that they have strong opinions on the search tools. During a development/implementation process these users – although most likely in minority – might be loud and demanding, since they know what to ask for. Organisations should keep in mind that these users are likely to represent only a small portion of the total user community. A strong voice should not allow them to marginalise the silent majority of less active searchers. However, since user groups B and C are small (and thus require fewer licenses) it may be feasible to buy them more advanced (and expensive) tools if they are considered important enough to the organisation.

Cluster F, finally – the knowledgeable users – is another small cluster, but what separates this segment from the rest is the fact that they wade through many result pages. They also spend a significant amount of time reading the documents they actually click on. If these are experienced searcher – and the use of many search terms suggest that they might be – they may also benefit from more sophisticated tools, and since they appear to be a small user group, this may not be too costly. As with clusters B and C, these users too may need tools that help

them visualise the search results in a non-linear way; automatically clustered or categorised according to some taxonomy. It has been shown that users are unable to effectively understand the content of a large data collection unless it can be visualised in ways that allows intuitive interaction with the data (Chang, Leggett, Furuta, Kerne, Williams, Burns & Bias, 2004). Tools to visualise clustered information could help these user groups form an overall understanding more efficiently.

Although some intranets may be very large both in terms of users and in content, they are all small compared to the public Web. It is therefore likely that technique not feasible for the public Web can successfully be applied to intranets. By realising this fact, and understanding that not all employees have the same search needs, behaviours or preferences, designers and developers of corporate intranets should be able to do better than provide one-size-fits-all out-of-the-box search solutions to their companies.

Limitations and Future work

The reason for clustering data in the first place is to reduce complicity so that patterns that would otherwise be hidden can surface. This reduction has the trade-off of losing details but this is a price that is acceptable given the increased understanding of the whole. One consequence of this in our work is that we have used average values in our vectors. Reporting that two users have the same average query length may hide the fact that the variance could differ significantly between the two. What we would interpret as similar behaviour could in fact be quite different, and this is a methodological limitation. Further, the same user can exhibit different usage patterns on different occasions, and our approach would not detect this. However, we think our approach is a reasonable approximation for the following reasons: Firstly, less than 30% of the users had logged more than one session. To the large majority, using averages had no negative effect. Secondly, it is also reasonable to assume that some of the users who did engage in multiple sessions actually behaved in a consistent way. To them, too, using averages did not affect the outcome negatively. Thirdly, using average values for the small group of remaining users who exhibited different search patterns during different sessions meant that their otherwise more distinct behaviours were somewhat flattened out and blurred. A possible result of this would be that the characteristics of the (small) clusters may not be as clear as they otherwise would have been, and the size might also have been slightly bigger. By and large, though, we think the clusters would have remained pretty much the same. This discussion shows, however, that deciding on a useful number of dimensions to include in the model is a delicate balancing act; a too complex model requires more processing power and may produce results that are difficult to interpret. We have suggested a level we believe is both feasible and useful but more research in this area is obviously needed.

To do an analysis of this kind, it is important to collect data from an extended time period so that users are given a chance to return. Chen & Cooper (2001) based their analysis on four weeks' worth of data. We have not had access to that amount of data but a week's worth of data is still significantly more than the single day analysis often used in studies of public web search engines. In addition, when comparing 7 days' worth of data to 25 days' worth of data (as we did in Stenmark & Jadaan, 2006), we saw only small changes; the users we studied were not very active. Another issue is that our analysis is based on users who actually did visit the search engine during the measured week; a large majority of the employees did not. To understand how they satisfied their information needs, other methods must be applied.

Although the clusters themselves are computer-generated in an automated fashion, the decisions regarding spatial layout that we have made have affected the result. When studying Figure 2 it becomes clear that adjacent cells can be quite similar whilst still being placed in different clusters. The decision where to draw the cluster borders may thus seem arbitrary, but, as explained earlier, these decisions were informed by analysing the topological error, the quantization error, and the Davies-Bouldin index to find the "best" places to draw the borders. The six clusters that emerge out of our work all have distinguishable centroids that have their own characteristic features. The interesting result is not whether we can find 4 or 8 clusters but the fact that there is more than one cluster, and we can identify them using this approach. This confirms previous findings that search engine users are not a homogeneous group of stereotypes that should be treated collectively. Although we expect similar results (i.e. half a dozen distinguishable clusters of users) to be found on other intranets, the exact number is likely to be context-specific and thus varies between organisations. The decision of how many clusters to opt for should therefore carefully be analysed before running the clustering algorithm chosen.

Other clustering algorithms than the one used by us may provide different results, and we invite more research in this area. However, the focus in this paper has not been to find and use the best or most efficient clustering algorithm, but to show the feasibility of using clustering techniques to identify different groups of user behaviour.

Conclusions

In this study, we have argued that there is no typical intranet searcher that behaves in one typical way and showed that it is problematic to have merely one search tool with one single interface. Using Self-Organising Maps, we have identified and described differences between segments of information seekers in intranets, and we can thus conclude that intranet search engine users are not a homogenous group. Instead, these users can be split up in segments, each with their particular behaviour characteristics.

In this particular study, we found six different clusters, but obviously a few of them predominate. The largest category, with nearly 80% of the users (consisting of clusters A, D and E), represents the "casual seekers". Many of these appear to be "fact seekers"; intranet users

looking for quickly retrieved answers. Useful and relevant descriptions and associated with the results links would probably help these users navigate more efficiently. This category is also likely to benefit from having precision boosted at the expense of recall.

The second category users, with 14% of the users (consisting of clusters B and C), apply a more holistic approach to information seeking and consequently have longer sessions, and more reading time. We suggest that these users would appreciate high recall and perhaps be willing to pay for this with lower precision. This should then be complemented with better search results visualisation that allowed for automatic clustering or categorisation.

The third and smallest category consists of cluster F users (7%). These are the information seeking savvy employees, most likely with both training and experience. They formulate longer queries and browse through more documents than do the other groups. Whether they prefer precision over recall or vice versa is unknown and may change from time to time.

We conclude that self-organising maps can successfully be used to find and identify clusters of intranet search engine usage behaviour using a standard combined log file, thus reducing the manual work required. We also conclude that such cluster can be used to better understand intranet users interaction with search engines thus help researchers and developers provide more targeted search solutions, instead of the current one-size-fits-all search engine.

Acknowledgments

This work was sponsored by Swedish Council for Working Life and Social Research via grant #2004-1268. The author is thankful to Henrik Strindberg who did a good and thorough job tuning the SOM toolbox parameters. Thanks are also due to the anonymous *JASIST* reviewers for their valuable comments that greatly improved this text.

References

Berry, M. J. & Linoff, G. (1997). Data mining techniques: For marketing, sales, and customer support. New York, NY: John Wiley & Sons.

Bilal, D. (2000). Children's use of the Yahooligans! Web search engine: I. Cognitive, physical, and affective behaviors on fact-based search tasks. Journal of the American Society for Information Science, 51(7), 646-665.

Bilal, D. (2001). Children's use of the Yahooligans! Web search engine: II. Cognitive, physical, and affective behaviors on research tasks. Journal of the American Society for Information Science, 52(2), 118–136

Buckland, M. & Gey, F. (1994). The relationship between Recall and Precision, Journal of the American Society for Information Science, 45(1), 12-19.

Chang, M., Leggett, J. J., Furuta, R., Kerne, A., Williams, J. P., Burns, S.A. & Bias, R.G. (2004) Collection understanding, Proceedings of JCDL 2004, Tuscon, AZ, June 7-11, 334-342

Chen, H.-M., & Cooper, M. D. (2001). Using clustering techniques to detect usage patterns in a web-based information system. Journal of the American Society for Information Science and Technology, 52(11), 888-904.

Chen, H., Houston, A.L., Sewell, R.R., & Schatz, B.R. (1998). Internet browsing and searching: User evaluations of category map and concept space techniques. Journal of the American Society for Information Science, 49(7), 582–603.

CIS (2008). SOM Toolbox homepage, Laboratory of Computer and Information Science (CIS), Department of Computer Science and Engineering, Helsinki University of Technology. URL: http://www.cis.hut.fi/projects/somtoolbox/ [April 2008]

Cooper, A. (1999). The inmates are running the asylum: Why high tech products drive us crazy and how to restore the sanity. Indianapolis: Sams.

Davies, D.L. & Bouldin D.W. (1979). A cluster separation measure, IEEE Transactions on Pattern Analysis and Machine Intelligence, 1(4), 224-227.

Desmet, H. (2001). Buying behavior study with basket analysis: pre-clustering with a Kohonen map, European Journal of Economic and Social Systems, 15(2), 17-30.

Eason, K., Richardson, S. & Yu, L. (2000). Patterns of use of electronic journals, Journal of Documentation, 49(4), 356-69.

Gevrey M., Worner S.P, Kasabov, N., Pitt, J. & Giraudel, J-L. (2006). Estimating Risk of Events Using SOM Models: A Case Study on Invasive Species Establishment. Ecological Modelling, 197(3-4), 361-372

Ghani, R. & Fano, A. (2002). Building recommender systems using a knowledge base of products semantics, in Proceedings of the Workshop on Recommendation and Personalization in E-Commerce, at the 2nd International on Adaptive Hypermedia and Adaptive Web Based Systems, Malaga, Spain.

Günter, S. & Burke, H. (2001). Validation indices for graph clustering, in Proceedings of the 3rd IAPR-TC15 Workshop on Graph-based Representations in Pattern Recognition, Ischia, Italy, 229–238.

Halkidi, M., Batistakis, Y. & Vazirgiannis, M. (2001) On Clustering Validation Techniques, Journal of Intelligent Information Systems, 17(2/3), 107–145.

He, D. & Göker, A. (2000). Detecting session boundaries from Web user logs, in Proceedings of the 22nd Annual Colloquium on Information Retrieval Research (ECIR), Sidney Sussex College, Cambridge, England.

Huang, X., Peng, F., An, A. and Schuurmans, D. (2004). Dynamic web log sessions identification with statistical language models, Journal of the American Society for Information Science and Technology, 55(13), 1290-1303.

Jansen, B. J. (2006). Search log analysis: What is it; what's been done; how to do it, Library and Information Science Research, 28(3), 407-432.

Jansen, B. & Spink, A. (2003). An Analysis of Web Documents Retrieved and Viewed, in Proceedings of ICIC 2003, Las Vegas, NE, 65-69.

Jansen, B., Spink, A., Bateman, J. & Saracevic, T. (1998). Real life information retrieval: A study of user queries on the web, ACM SIGIR Forum, 32(1), 5-17.

Kotler, P., Armstrong, G., Cunningham, M. H. & Warren R. (1996). Principles of Marketing, 7th edition. Prentice-Hall.

Li, H., Cao, Y., Xu, J., Hu, Y., Li, S. & Meyerzon, D. (2005). A new approach to intranet search based on information extraction, in Proceedings of CIKM '05, Bremen, Germany, Oct. 31-Nov.5, 460-468.

Liu, F., Yu, C. & Meng, W. (2002). Personalized web search by mapping user queries to categories, in Proceedings of CIKM '02, McLean, VA, Nov. 4-9, 558-565.

Machón, I. & López, H. (2006). End-point detection of the aerobic phase in a biological reactor using SOM and clustering algorithms. Engineering Applications of Artificial Intelligence, 19(1), 19–28.

Maarek, Y.S. & Ben-Shaul, I.Z. (1996). Automatically organizing bookmarks per contents, Computer Networks and ISDN Systems, 28, 1321-1333.

Moore, J. L., Erdelez, S. & He, W. (2007). The search experience variable in information behavior research, Journal of the American Society for Information Science and Technology, 58(19), 1529-1546.

Nielsen, J. (1999). User Interface Directions for the Web, Communications of the ACM, 42(1), 65-72.

Silverstein C., Henzinger, M., Marais, H. & Moricz, M. (1998). Analysis of a Very Large AltaVista Query Log, Digital SRC Technical Note #1998-014, October 26.

Spink, A. & Jansen, B. (2004). Web Search: Public searching of the web. Kluwer Academic Publisher.

Stenmark, D. (2005). One week with a corporate search engine: A time-based analysis of intranet information seeking, in Proceedings of AMCIS 2005, Omaha, Nebraska, August 11-14, 2306-2316.

Stenmark, D. & Jadaan, T. (2006). Intranet Users' Information-Seeking Behaviour: A Longitudinal Study of Search Engine Logs, in Proceedings of ASIS&T 2006, Austin, Texas, November 3-8.

Strindberg, H. (2006). Mining a corporate intranet for user segments of information seeking behavior. Master Thesis in Informatics, University of Gothenburg, Sweden.

Vesanto, J. & Alhoniemi, E. (2000). Clustering of the self-organizing map, IEEE transactions on neural networks, 11(3), 586-600.

Vesanto, J., Himberg, J., Alhoniemi, E. & Parhankangas, J. (1999). Self-organizing map in matlab: the SOM toolbox, in Proceedings of the Matlab DSP Conference '99, Espoo, Finland, November 16-17, 35-40.

Wang, L., Jiang, M., Lu, Y., Noe, F. & Smith, J. C. (2006). Self-Organizing Map Clustering Analysis for Molecular Data, in Proceedings of Third International Symposium on Neural Networks, Chengdu, China, May 28-June 1, 1250-1255.

Xu, R. and Wunsch, D. (2005). Survey of clustering algorithms. IEEE Transactions on Neural Networks, 13(3), 645–678.

Appendix

Projection considerations

When transforming from a high-dimension input such as 7,902 11-dimensional vectors to a low-dimension output such as a 50 neuron map, two types of errors are introduced; the quantization error and the topological error (Desmet, 2001). The quantization error (QE) is the average distance between the input layer vector and the neuron. The topological error (TE) measures the proportion of vectors for which the best matching unit, i.e., the neuron closest to the input space, is not closely related in the output layer

Desmet (2001) suggests that the choice of dimensions for the output layer of SOM has implications on the quality of the projection and has to be examined separately, as it is data dependent. We therefore calculated the QE and the TE for fifteen different configurations (the x-axis in the diagrams in Figure X). The first two diagrams show how the Y and X dimensions varied, respectively. The product of X×Y is the number of neurons in the configuration and shown in the fifth and last diagram.

As Figure X shows, the QE (third diagram from the top) decreases steadily with the growing number of neurons while the TE (third diagram from the top) fluctuates. Fifty neurons (i.e., 5×10 as in configurations 5 or 10×5 as in configuration 6) appeared to be a good trade-off between computational efforts required on the one hand and the quality and visualibility of the output on the other hand. Comparing configurations 5 and 6, we see that a 10×5 sheet (configurations 6) is to prefer over a 5×10 sheet (configurations 5) since it has a lower TE value while the QE remains constant (Figure A).

Cluster detection and validation

Automatic clustering is an unsupervised method and therefore there is no knowing in before-hand how many clusters will be produced. What we do know is that creating artificial borders between neurons will introduce errors and to asses these, some kind of external clustering results validation has to be applied (Halkidi, Batistakis & Vazirgiannis, 2001).

Günter and Burke (2001) point out that when the "correct" number of clusters is not known, one can execute a clustering algorithm multiple times, varying the number of clusters in each run from some minimum to

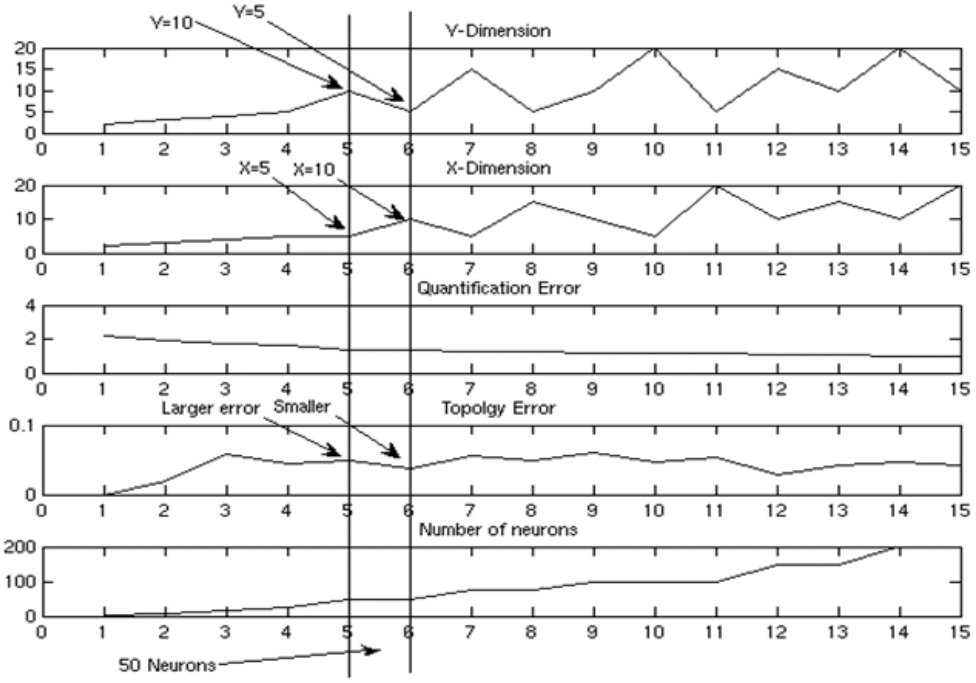

FIG. A. Coordinated plot of X-dimensions, Y-dimensions, TE, QE and number of neurons.

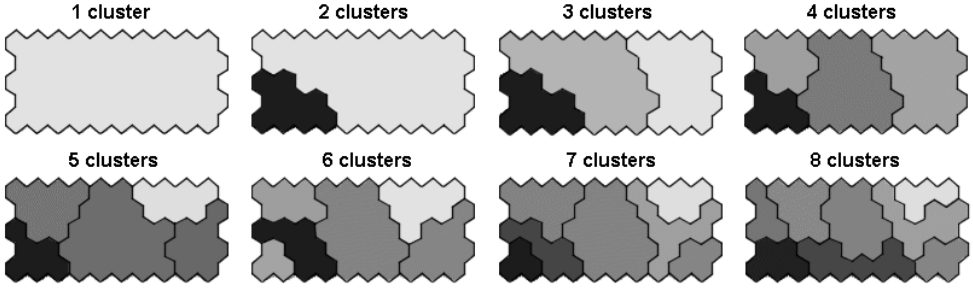

FIG. B. SOM-generated clusters on a 10 × 5 neuron sheet

some maximum value. Since neither too few nor too many clusters would be useful from an analytic point of view, we therefore used SOM's k-mean clustering feature to generate up to eight clusters (see Figure B).

For each configuration obtained a validation index was computed. Günter and Burke (2001) define a cluster validation index as a number that indicates the quality of a given clustering. The SOM toolkit has a number of built-in assessment functions and we applied the Davies-Bouldin (1979) index since it is a well-known index used in many SOM applications in a variety of different fields (e.g., Gevrey, Worner, Kasabov, Pitt & Giraudel, 2006; Machón & López, 2006; Wang, Jiang, Lu, Noe & Smith, 2006). Plotting the amount of error as a function of the number of clusters, the 6 clusters configuration yielded the best index value and was thus selected (see Figure C).

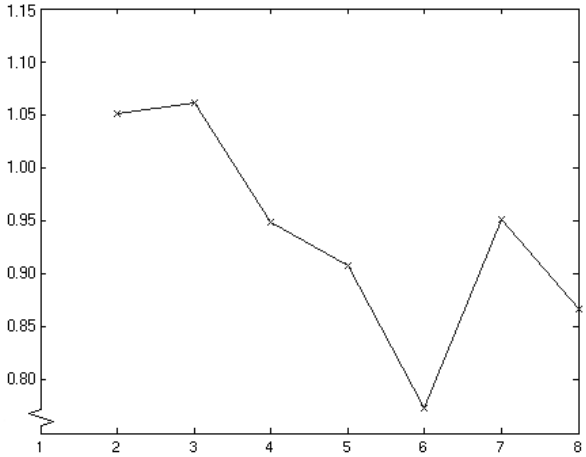

FIG. C. Movement of the Davies & Bouldin (1979) index. The x-axis denotes the number of clusters. Note the local minimum at six clusters.