Newsgroup Clustering Based On User Behavior — A Recommendation Algebra

Jussi Karlgren
jussi@sics.se

March 1994 *

Abstract

User models are a tool for guiding system behavior in interactive systems, and their utility and properties, desirable and undesirable, have been investigated in this context. There are several ways of utilizing information about the user that have *not* been implemented, however. In this paper a scheme for users to peek at other users' user models to extract information is proposed, in an information retrieval or information filtering domain. The material used for the study is a set of .newsrc files.

Keywords

Data extraction from user models; clustering.

Background

Compared to an ordinarily untidy bookcase computerized systems for information retrieval may not always be better. In a normal bookcase interesting documents may be found next to each other, and someone looking for a certain document may unexpectedly find other interesting documents in the vicinity. They are interesting because someone placed them there, and they are placed there because they have some relation to the original document. An unorganized bookcase will self-organize – somewhat unsystematically – based on the user's behavior. In fact, in a library or a bookstore, people around an interesting bookcase tend to be interesting people. You tend to be able to get good reading

*This work has been made possible by the generous IBM Electrum Scholarship 1992.

tips from them. Similarly, a good librarian will remember that a certain book tends to be read by a certain set of people, and another book by the same set of people, and that there may be a similarity between the books, even though they may not be catalogized together – as of course, they often will not be. Anyone who has tried to organize a bookcase by topic knows how many cases of unexpected category conflict one encounters.

The situation and the request in real information retrieval situations can often be formulated as a form of "I read *A Good Book* – I want more of the same"[1], posed to a librarian or to a colleague – or to a number of colleagues. The idea is, as in all document classification, to use a distance measure to build a document space, and to use clustering algorithms to categorize documents in the space. This is a standard method, and most often, the distance measure used has to do with the content of the document, as judged by an intermediary such as a librarian or a documentalist, or by a text search system, using keywords or free-text search.

In the application outlined in this article, the distance measure will be based on knowledge about the *user* or knowledge about the *use* of the document rather than knowledge about the *content* or *genre* of the document. This knowledge is extracted from user models that indicate the preference of users.

The Domain

Initially, an experiment was performed where novels and video rental movies were used as a document base, but when the data matrix proved to be too sparse, newsgroups on the Usenet News discussion and information dissemination network were used instead. The data used were .newsrc files of Usenet News users. The results will generalize in an obvious way from "documents" that in fact are Usenet News newsgroups to other documents, and further to other similar retrievable objects, information sources, or even patterns of behavior.

User Models

User modeling has found numerous applications and involves numerous research projects, usually having to do with tailoring output from interactive systems to suit the needs of particular users or user types. This paper will assume the existence of a very simple user model, essentially containing user grades on documents.

One of the central questions when using user modeling techniques is how the contents of the user model – the grades, in our case – are gathered. This is an interesting question in itself, but will be left aside in this paper. Grades can conceivably be collected through explicit questioning of the user (Rich, 1979)

[1] Or, as put by Benny Brodda (1992): "Gimme More Of That".

through relevance feedback of some sort (Salton & McGill, 1983) or through inference on user behavior (Kass, 1991). In the application we have in mind now, whether the grades are obtained by explicit user recommendation or user behavior, either by querying users on their assessments of document relevance, or by examining user behavior to determine which documents actually are accessed, the methods for *using* the information in them will be similar. The system will gather information about several users to service the needs of one specific user – in effect constructing a dynamic *user stereotype* for a specific query.

An Algorithm for Computing Proximity

The information in a set of simple user models can be used to compute a proximity measure between documents. The first step is to define *interest* as a relation between a user and a document. The user model assumed in this paper will be very simple, as noted in the previous section, and will be a simple collection of grades:

```
grade(User,Document,Grade).
```

The user model contains a vector of user grades. The documents in the document base are graded by a user to be good "`+`", bad "`-`", or not accessed "`0`". The user X is *interested* in a document if the grade of that document is "`+`". The user X is *uninterested* in a document if the grade of that document is "`-`". The user X does *not know* the document if the grade of that document is "`0`"

A more complex grade palette will naturally need further formalization: in this case, all we need is a starting point to test the validity of the model.

Interest As A Proximity Measure

The basis for the algorithm is the following statement, which we will call the **Recommendation Hypothesis**: If a user A is interested in documents K and L, and another user X is interested in K, it is likely that X will also like L.

Phrased differently the hypothesis will be: If users agree on a document they will agree on others as well.

The starting point for the discussion was to find an adequate reply to the request "I read *A Good Book* – I want more of the same". We will now inspect what cases may occur if similarity is defined according to the Recommendation Hypothesis. User X poses the request above to the database. The database contains, among others, entries for the users A, B, and C as in the table below.

User X will need to figure out what grades to pay most attention to. It is obvious that user A has most to say: users X and A agree on the quality of *A Good Book*. User A also has an interest in document K. It seems reasonable to assume that document K may be interesting for user X as well. It is equally

User	*A Good Book*	*K*	*L*	*M*
A	+	+	-	0
B	-	+	-	0
C	0	+	-	0

Table 1: Sample User Models

$\otimes$	+	-	0
+	Another good book.	Warning.	Don't know.
-	A better book.	Another uninteresting book.	Don't know.
0	Don't know.	Don't know.	Don't know.

Table 2: A Qualitative Recommendation Algebra

reasonable to assume that user *C*, who has not read *A Good Book*, will have little to contribute in this discussion, and that document *M*, of which no user can say much will have an indeterminate grading as a result of the models inspected. Formalizing these observations, we will define *recommendations* as a product of interests.

Recommendations – An Algebra on Interests

Recommendations will be defined as products of interests. Given *A Good Book*, the parameters are what interest the user shows in it, and what the user can say of other books. The matrix in table **??** covers the cases that can arise. The leftmost vertical column indicates the grades the user has given *A Good Book*, and the top row, the grades a user has given other books.

Now, having defined recommendations, the question is how to use it. The likelihood of a recommendations being useful grows with the number of users that agree, and the number of documents they agree on. The whole point is to sum recommendations over all users. To do this, we need to quantify them. The discussion in the preceding section has defined the matrix shown in table **??** for the recommendation operator $\otimes$.

The question is what values to insert in the matrix in the cells now containing question marks. The simplest alternative is the one shown in table **??**.

Another realization of $\otimes$ is represented by the matrix in table **??**.

We will in the experiment section below concentrate on the simpler algebra defined in table **??**, and defer further discussion of more complex algebras.

$\otimes$	+	-	0
+	1	$?_1$	0
-	$?_2$	$?_3$	0
0	0	0	0

Table 3: A Template For A Quantitative Recommendation Algebra

$\otimes$	+	-	0
+	1	0	0
-	0	0	0
0	0	0	0

Table 4: A Quantitative Recommendation Algebra

$\otimes$	+	-	0
+	1	-1	0
-	-0,5	0,5	0
0	0	0	0

Table 5: Another Quantitative Recommendation Algebra

Proximity – A Sum Over Recommendations

The proximity from a document to another – from *A Good Book* to document K – can be defined as a sum over all readers' interests in *A Good Book* and K:

$$proximity(document_{AGB}, document_K) =$$

$$\sum_i (interest(reader_i, document_{AGB}) \otimes interest(reader_i, document_K))$$

This sum will then be used for clustering documents, using any standard algorithm. Note, however, that as we have defined them, the proximity measure does not need to be symmetrical: the proximity from *A Good Book* to document K does not have to be equal to the proximity from document K to *A Good Book*. This naturally depends on the definition of $\otimes$. The two algebras defined in tables **??** and **??** are different in this respect: the first is symmetric while the second is not. Standard clustering algorithms require the proximity or distance measure to be symmetric: this is one reason we here only address such simpler algebras.

Larger seed set

A natural extension of the discussion so far is to use several documents as a starting point – a *seed set* – for the query: "I read *These Good Books* – I want more of the same". It is not intuitively clear how the grades for one document should be compiled into one grade: a simple sum will probably not give enough credit to documents that cooccur with a large part of the seed set. We have not addressed this problem in this first study.

Clustering

As noted above, standard clustering algorithms from statistical literature, the proximity measure has to be symmetrical. (See for instance Miyamoto, 1989). This is not necessary for any reason inherent in the nature of proximity or of documents, but for practical reasons, to be able to use standard, unmodified clustering algorithms, we have chosen to use the algebra defined in table **??**. The clustering algorithm we use is a standard average linkage agglomerative method.

Experiments

Data Failure: Pilot Experiment

25 subjects were asked to grade 150 novels and 150 video movies with one of the three grades above. This material was processed with different $\otimes$ matrices.

```
news.announce.conferences: 1-5260
news.announce.important: 1-51
alt.tv.rockford-files: 1-253
swnet.jobs: 1-232
misc.jobs.offered! 1-31344
dk.jobs! 1-24
comp.lang.prolog: 1-9033
comp.risks: 1-6314
comp.society.futures! 1-3343
comp.society.privacy: 1-1993
comp.cog-eng! 1-2247
comp.ai.nlang-know-rep: 1-1577
comp.ai! 1-17353
sics.general: 1-244
sics.sicstus: 1-404
sics.syschanges: 1-1060
sics.personal.forening: 1-76
```

Table 6: An excerpt from a typical .newsrc file

This data proved unsatisfactory – the matrix was simply too sparse. The data set was too heterogenous, and the reasons behind reading novels and watchin video movies too diverse, even in a the relatively homogenous population the subjects were taken from: all were graduate students of Columbia University.

Usenet News Domain

A number of .newsrc files were gathered and processed. A typical .newsrc file has an appearance as in the excerpt shown in table **??**. The newsgroup name is followed by a character ":" or "!" which indicates if the newsgroup is subscribed to or not. The digits following the subscription character indicate which messages in the newsgroup have been read. These could be utilized to refine the grade set from binary to a continuous scale, but we have elected to keep the model simple, initially. The $\otimes$ matrix used was the one shown in table **??**. One of the good points of using Usenet News data is that it is reasonably easy to validate the clustering by the newsgroup names, which are fairly indicative of content.

$\otimes$	:	!
:	1	0
!	0	0

Table 7: Experiment Quantitative Recommendation Algebra

Experiment 1: n = 60

In the first experiment sixty .newsrc files were used. In this data set, the fifty most subscribed[2] newsgroups are the ones shown in table **??**. If recommendation statistics for these newsgroups are calculated we will get the matrix shown in table **??**. This data is transformed to a distance measure matrix to get the matrix in table **??** by adding one to each element (to avoid zero values) and then inverting them all.

The distance matrix in table **??** clustered as shown in table **??**. Some of the clusters are remarkably well put together, whereas others may be seem more haphazard. There is a local news set (including job ads!), a local system set, a technical set, a Prolog set, an AI set, a swnet set, a erotic picture set, and a couple of less easily labelable but nonetheless not completely weird sets.

Experiment 2: n = 600

In the second experiment six hundred .newsrc files were used. In this data set, the fifty most subscribed newsgroups are the ones shown in table **??**. The recommendation statistics for these newsgroups are shown in table **??**. This data is again, as in the previous experiment transformed to a distance measure matrix to get the matrix in table **??** by adding one to each element (to avoid zero values) and then inverting them all.

The distance matrix in table **??** clustered as shown in table **??**. Again, some of the clusters are remarkably well put together, whereas others may be seem more haphazard. Most of them have a reasonably well-defined profile, however. There is a general Swedish discussion group set, a programming group set, a network set, a picture set, a Uppsala news set (including classified ads and job announcements), and finally, the two local newsgroups.

[2] It would be more interesting to be able to say that these are the most *read* newsgroups. This we cannot do, although it is conceivable that we could extract this from the numerical info on *seen* articles that follows the subscription information. It is not self-evident, however, that an article which is *seen* has also been *read*. For this reason, initially, we will be cautious about making the inference.

local.general	swnet.sources.list
swnet.jobs	swnet.diverse
local.syschanges	kth.sun
local.personal.forening	kth.mac
local.alla	kth.general
local.sicstus	eunet.news
comp.lang.prolog	comp.ai.vision
local.protokoll	comp.ai.nlang-know-rep
local.mac	comp.ai.edu
local.system	swnet.mail.map
local.library	sci.logic
comp.ai	local.ai-in-medicine
news.announce.conferences	alt.binaries.pictures.supermodels
local.test	alt.binaries.pictures.erotica.blondes
news.announce.important	alt.binaries.pictures.erotica.female
swnet.pryltorg	alt.binaries.pictures.erotica
swnet.ai.neural-nets	news.answers
kth.unix	comp.cog-eng
comp.ai.shells	swnet.conferences
news.announce.newusers	aus.jobs
misc.jobs.offered	comp.newprod
comp.ai.digest	swnet.sunet-info
kth.data	swnet.org.snus
swnet.utbildning.grundbulten	alt.crackers
swnet.sys.cv	comp.sys.workstations

Table 8: The fifty most popular newsgroups in experiment 1 in descending order of popularity.

```
16, 9, 9,11,10, 8, 7,10, 9, 7, 9, 7
 9,13, 7, 6, 5, 5, 7, 5, 4, 5, 5, 6
 9, 7,11, 8, 5, 7, 7, 5, 6, 6, 5, 5
11, 6, 8,11, 8, 7, 5, 9, 8, 6, 8, 4
10, 5, 5, 8,11, 8, 5, 7, 6, 6, 5, 3 .
 8, 5, 7, 7, 8,10, 6, 5, 5, 7, 4, 3 .
 7, 7, 7, 5, 5, 6,10, 3, 4, 4, 4, 5 .
10, 5, 5, 9, 7, 5, 3,12, 7, 5, 9, 4
 9, 4, 6, 8, 6, 5, 4, 7, 9, 4, 7, 4
 7, 5, 6, 6, 6, 7, 4, 5, 4, 9, 4, 2
 9, 5, 5, 8, 5, 4, 4, 9, 7, 4, 9, 5
 7, 6, 5, 4, 3, 3, 5, 4, 4, 2, 5, 8
                ...
```

Table 9: Excerpt from proximity matrix for the fifty most popular newsgroups inexperiment 1.

0.05,	0.10,	0.10,	0.08,	0.09,	0.11,	0.12,	0.09,	0.10,	0.12,	0.10,	0.12
0.10,	0.07,	0.12,	0.14,	0.16,	0.16,	0.12,	0.16,	0.20,	0.16,	0.16,	0.14
0.10,	0.12,	0.08,	0.11,	0.16,	0.12,	0.12,	0.16,	0.14,	0.14,	0.16,	0.16
0.08,	0.14,	0.11,	0.08,	0.11,	0.12,	0.16,	0.10,	0.11,	0.14,	0.11,	0.20
0.09,	0.16,	0.16,	0.11,	0.08,	0.11,	0.16,	0.12,	0.14,	0.14,	0.16,	0.25
0.11,	0.16,	0.12,	0.12,	0.11,	0.09,	0.14,	0.16,	0.16,	0.12,	0.20,	0.25
0.12,	0.12,	0.12,	0.16,	0.16,	0.14,	0.09,	0.25,	0.20,	0.20,	0.20,	0.16
0.09,	0.16,	0.16,	0.10,	0.12,	0.16,	0.25,	0.08,	0.12,	0.16,	0.10,	0.20
0.10,	0.20,	0.14,	0.11,	0.14,	0.16,	0.20,	0.12,	0.10,	0.20,	0.12,	0.20
0.12,	0.16,	0.14,	0.14,	0.14,	0.12,	0.20,	0.16,	0.20,	0.10,	0.20,	0.33
0.10,	0.16,	0.16,	0.11,	0.16,	0.20,	0.20,	0.10,	0.12,	0.20,	0.10,	0.16
0.12,	0.14,	0.16,	0.20,	0.25,	0.25,	0.16,	0.20,	0.20,	0.33,	0.16,	0.11
					...						

Table 10: Excerpt from distance matrix for the fifty most popular newsgroups in experiment 1.

Relevance as a Tool for Information Retrieval

A distance or proximity measure based on heuristics like these can only be expected to produce useful results up to a point. This measure is in a certain sense orthogonal to standard metrics for document classification: indexing schemes, word statistics, and the like. It does not analyze the material with respect to its *content*, as do standard information retrieval metrics. Neither does it take into account the *formal guise* of the material, as do some experimental metrics, for instance in calculating probable text *genre* (Karlgren & Cutting, 1994). This means that it can be expected to produce results as an additional layer added on to existing and future traditional content- and genre-based information retrieval mechanisms, not as a complete tool on its own.

Relevance to the Study of User Modeling

We do not expect the tool to be very useful in the Usenet News domain: this experiment was just to show the utility the algorithm on easily available data. In the Usenet News domain, most users have a reasonable overview of what is available on the net. However, we will evaluate the technique in the IntFilter project at SICS and Stockholm university. The IntFilter project aims at producing interactive filtering tools for information flows, and we will use the clustering mechanism described here to produce standard stereotypical .newsrc user models for new users. New users can then get a whole package of files to start the interaction with, without having to immerse themselves in the entire Usenet News flow.

local.general swnet.jobs local.syschanges local.personal.forening	local.protokoll local.mac local.library news.announce.important
swnet.utbildning.grundbulten swnet.sys.cv swnet.sources.list swnet.diverse swnet.conferences	univ.unix univ.mac univ.general eunet.news swnet.mail.map
comp.ai news.announce.conferences comp.ai.shells comp.ai.nlang-know-rep	local.system local.test univ.data univ.sun
swnet.ai.neural-nets comp.ai.edu	alt.crackers comp.sys.workstations
comp.ai.vision sci.logic	news.answers news.announce.newusers
swnet.sunet-info swnet.org.snus	swnet.pryltorg aus.jobs
alt.binaries.pictures.supermodels alt.binaries.pictures.erotica.blondes alt.binaries.pictures.erotica.female alt.binaries.pictures.erotica	local.alla local.sicstus comp.lang.prolog
misc.jobs.offered	comp.ai.digest
local.ai-in-medicine	comp.cog-eng
comp.newprod	

Table 11: Clusters in experiment 1.

news.announce.newusers	alt.sex.stories
swnet.jobs	swnet.ai.neural-nets
swnet.pryltorg	uppsala.mail
uppsala.news	swnet.sunet-info
uppsala.general	swnet.test
swnet.general	alt.binaries.pictures.utilities
swnet.wanted	swnet.svenska
swnet.diverse	nordunet.general
swnet.conferences	alt.sex
uppsala.test	alt.3d
uppsala.games	swnet.siren
local.news	swnet.sys.sun
swnet.unix	swnet.sources
swnet.sys.amiga	comp.sources.unix
swnet.sys.ibm.pc	alt.sex.pictures.female
swnet.followup	swnet.snus
swnet.sys.mac	alt.sources
swnet.mail	gnu.announce
alt.binaries.pictures.supermodels	comp.sources.x
comp.lang.c++	swnet.sources.list
swnet.politik	alt.binaries.pictures.misc
news.announce.important	comp.lang.c
alt.binaries.pictures	alt.binaries.pictures.erotica
local.diverse	swnet.sys.sun.flash
comp.lang.prolog	swnet.mac

Table 12: The fifty most popular newsgroups in experiment 2 in descending order of popularity.

```
209, 66, 59, 48, 56, 57, 50, 45, 50, 57, 40, 24, 47
 66,161,115, 77, 79, 85, 85, 63, 79, 63, 58, 45, 71
 59,115,144, 74, 77, 80, 84, 63, 74, 66, 53, 38, 70
 48, 77, 74,119, 86, 51, 45, 39, 57, 45, 72, 48, 45
 56, 79, 77, 86,115, 63, 52, 42, 59, 59, 68, 40, 54
 57, 85, 80, 51, 63, 95, 69, 54, 65, 62, 40, 22, 65   .
 50, 85, 84, 45, 52, 69, 92, 42, 59, 55, 36, 23, 61   .
 45, 63, 63, 39, 42, 54, 42, 90, 40, 42, 24, 28, 40   .
 50, 79, 74, 57, 59, 65, 59, 40, 90, 55, 37, 27, 57
 57, 63, 66, 45, 59, 62, 55, 42, 55, 87, 33, 17, 46
 40, 58, 53, 72, 68, 40, 36, 24, 37, 33, 86, 34, 37
 24, 45, 38, 48, 40, 22, 23, 28, 27, 17, 34, 84, 22
 47, 71, 70, 45, 54, 65, 61, 40, 57, 46, 37, 22, 83
                 ...
```

Table 13: Excerpt from proximity matrix for the fifty most popular newsgroups inexperiment 2.

```
0.004761904761904762, 0.014925373134328358, 0.016666666666666666, ...
0.014925373134328358, 0.006172839506172839, 0.008620689655172414, ...
0.016666666666666666, 0.008620689655172414, 0.006896551724137931, ...
0.020408163265306120, 0.012820512820512820, 0.013333333333333334, ...
                        ...
```

Table 14: Excerpt from distance matrix for the fifty most popular newsgroups in experiment 2.

swnet.wanted	comp.lang.c++
swnet.conferences	comp.lang.prolog
uppsala.test	alt.sex
swnet.unix	comp.sources.unix
swnet.sys.amiga	alt.sources
swnet.sys.ibm.pc	gnu.announce
swnet.followup	comp.sources.x
swnet.sys.mac	swnet.sources.list
swnet.mail	comp.lang.c
swnet.politik	
news.announce.important	swnet.ai.neural-nets
swnet.sunet-info	swnet.test
swnet.sys.sun	nordunet.general
swnet.sources	swnet.siren
swnet.snus	
news.announce.newusers	alt.binaries.pictures
swnet.jobs	alt.binaries.pictures.utilities
swnet.pryltorg	alt.sex.stories
uppsala.news	
uppsala.general	local.news
swnet.general	local.diverse
alt.3d	
swnet.sys.sun.flash	
swnet.diverse	uppsala.games
uppsala.mail	alt.binaries.pictures.supermodels
swnet.svenska	alt.sex.pictures.female
alt.binaries.pictures.misc	alt.binaries.pictures.erotica

Table 15: Clusters in experiment 2.

Interactivity Aspects

How to design an interaction using techniques such as these, that by necessity will seem complex to the casual user is a question we address in a separate publication (Karlgren *et al.*, 1994); in connection with this technique we must note that a complex algebra, such as the one tentatively defined in table **??** will be difficult to explain to the user. Indeed, one of the reasons we did not pursue the study of it further was the complexity involved in debugging output from the program. Even in development stages, when the algorithm was parametrizable and highly salient, its behavior became complex. This is not the last word on algebra design, however: the alternative quantifications must be studied further before judgement is passed in this matter.

Integrity Aspects

An obvious stumbling block for utilizing user models in this manner is that of *user integrity.* Integrity questions are important to consider, and difficult to resolve in a straightforward way. Indeed, this experiment itself is an illustration of the personal integrity problem complex. The .newsrc files used in the experiment were taken from open available systems at universities and research institutes the IntFilter project has access to. The users themselves were not asked, but were assumed to have given their permission implicitly by their having the protection of the .newsrc files set so that they were readable outside their immediate work group. The .newsrc files were immediately de-identified, so none of the data can be attributed to any single user – however, no users actually were made aware of the fact that their reading habits were bandied about for public scrutiny. A tentative solution to empower the user, would be to allow the user an unlimited number of identities, thus letting users partition their reading habits by *pseudonyms* (Bratt *et al.*, 1983). Obviously this does not solve all problems, but a solution of this type at least redresses some of the balance that the system otherwise takes from the user.

References

Hans Iwan Bratt, Hans Karlgren, Ulf Keijer, Tomas Ohlin, Gunnar Rodin. (1983). "En liberal datapolitik", *Tempus*, 16-19/12, Stockholm.

Benny Brodda. (1992). "Gimmie More O'That", *IRI Publication*, Stockholm: IRI.

Jussi Karlgren and Douglass Cutting. 1994. "Recognizing Text Genres with Simple Metrics Using Discriminant Analysis" Submitted to Coling 1994, Kyoto.

Jussi Karlgren, Kristina Höök, Ann Lantz, Jacob Palme, Daniel Pargman. (1994). "The glass box user model for filtering." Submitted to the First International Conference On User Modeling, Cape Cod.

Jussi Karlgren . (1992). *The Interaction of Discourse Modality and User Expectations in Human-Computer Dialog*, Licentiate Thesis at the Department of Computer and Systems Sciences, University of Stockholm.

Jussi Karlgren. (1990). "An Algebra for Recommendations", *Syslab Working Paper* 179, Department of Computer and System Sciences, Stockholm University, Stockholm.

Robert Kass. (1991). "Building a User Model Implicitly from a Cooperative Advisory Dialog", UMUAI 1:3, pp.203-258

Sadaako Miyamoto. (1989). *Fuzzy Sets in Information Retrieval and Cluster Analysis*. Dordrecht: Kluwer.

Elaine Rich. (1979). "User Modeling via Stereotypes" *Cognitive Science*, vol 3:pp 329-354.

Gerald Salton and Michael McGill. (1983). *Introduction to Modern Information Retrieval* New York: McGraw-Hill.