# Licentiate Thesis Proposal
# Clustering and Case-Based Reasoning for User Stereotypes

Mikael Sollenborn
Department of Computer Science and Engineering
Mälardalen University
Västerås, Sweden
mikael.sollenborn@mdh.se

## Abstract

This document presents a proposal for the contents of a licentiate thesis in computer science at Mälardalen University, Sweden. The main subject of the thesis is the usage and creation of user stereotypes, aided by the use of clustering techniques to find similar groups of users. To reason about and maintain the user stereotypes, we primarily use Case-Based Reasoning (CBR). The techniques presented will be used in two seemingly separate, yet somewhat related, application domains: web filtering, and medical diagnosis[1].

Supervisors:

- Docent Peter Funk
  Department of Computer Science and Engineering
  Mälardalen University

- Professor Björn Lisper
  Department of Computer Science and Engineering
  Mälardalen University

- Dr Bo von Schéele
  The Institute for Psychosocial Medicine
  Karolinska Institutet

[1] The research is funded by KK-Stiftelsen, PBM-StressMedicine AB and Mälardalen University.

## 1. Introduction

In many situations there is an advantage in making the assumption that, despite the individuality of every person, similar behavioural patterns and characteristics can be extracted by studying a population. By doing this, people can be classified into groups. There is, however, a difference between a group as such, and (user) *stereotypes,* which will be the main focus of this licentiate proposal and the following thesis. A stereotype is a representation of a group of people as a single individual, which exists not as a real person, but only as an extraction of the most common features of a group of people. The term “user” can imply several things: it could mean an active user, such as a person browsing the contents of a web site, or a passive user, such as a “user” of treatments as proposed by a physician. We will occasionally use alternative terms preceding “stereotype” to be more precise about what particular kind of user stereotype we are looking at, e.g. patient stereotypes. Depending on the application and the amount of data available, the construction of stereotypes can be done manually or by automatic extraction of similar features, or any combination thereof. The focus in the proposed licentiate thesis will be on automatic creation and identification of stereotypes using clustering techniques.

The secondary focus of the licentiate thesis is on *Case-Based Reasoning* (CBR). Looking at each user stereotype as a case in a case library, it becomes natural to use CBR as the primary tool for reasoning about and managing stereotype. We will refer to the case representation of a user stereotype as a *stereotype case*.

The licentiate thesis will illustrate the usefulness of the combined approach of user stereotypes, clustering, and Case-Based Reasoning, by looking at distinctly separate application domains. Although this separation was partly accidental (due to the liquidation of the main participating company, which led to collaborations with another company in a different domain), the change of application domain actually helped to generalise the research further and shows the versatility of the proposed approach (more about this and the separate projects below).

The paper is organized as follows. The following section gives a background to the most important terms and methods being used in the licentiate thesis work. Section 3 deals with the motivation behind this work. Section 4 covers related work, and section 5 specifies the contributions made by the author to this field of research. In section 6, there is a thesis outline, followed by a time plan for the licentiate thesis in section 7. The last section looks into future work, looking beyond the licentiate thesis.

## 2 Background

This section covers the most important terms and methods used throughout the licentiate thesis proposal.

### 2.1 Case-Based Reasoning

Case-Based Reasoning (CBR) is both a model of human reasoning, and a method used to create “intelligent” systems. As a model, CBR is based on a number of key observations. The first observation is the fact that most of the problems a decision maker has to handle aren’t unique. When encountered with a new problem, novices and experts often reason by analogy, comparing the current situation with earlier problems encountered. The second observation is that when solving new problems, people typically reuse solutions from similar problems, adapting the solution to suit the current circumstances. In summary, the CBR model of human reasoning suggest that people reason by analogy, remembering past experiences.

CBR is a method for building intelligent systems based on reuse of past cases. Building a case-library covering the area in question is essential. The case library needs to cover a sufficiently large part of the problem space from the start, as adaptations to new problems are often hard to make if there are no stored cases similar enough to the new problem.

A case typically consists of a problem description, a set of identifying features, and a solution to the problem.

The CBR problem solving cycle is often referred to as the 4 RE:s: REtrieve, REuse, REvise, REtain, as illustrated in Figure 1.

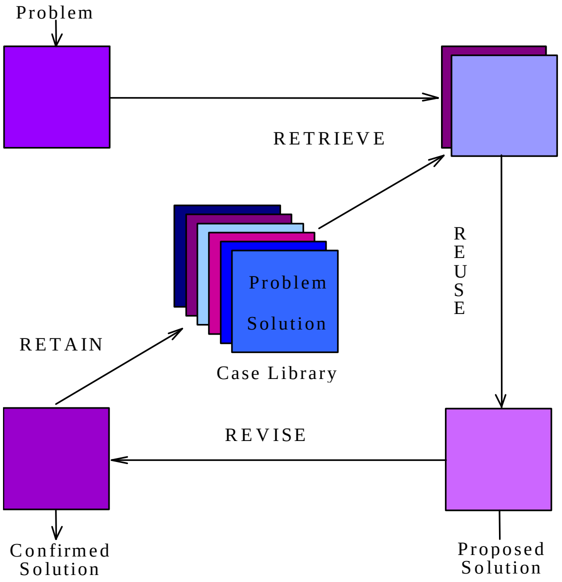

Figure 1. The CBR problem solving cycle

In the first step, Retrieve, the cases most similar to the current problem are selected using some kind of similarity metric, such as Nearest Neighbor or Inductive Retrieval.

In the Reuse step, the most similar of the cases selected in the first step is determined using additional similarity reasoning. If the current problem and the closest matching case are still dissimilar, the solution to the closest matching case is adapted using domain-specific rules. A proposed solution is then presented to the system user.

If the suggested solution was inappropriate, a Revision has to be made, based on the error report, which may be manual or automatically inferred. The confirmed solution is then presented.

In the last step, if the problem and/or the solution differed substantially from the closest case, the problem along with the solution is Retained in the case-base for later use [WATSON1997] [KOLOD1993].

### 2.2 User Stereotypes

A user model represents the current knowledge about a user as an individual, gathered through measurements, questionnaires, observation etc.

A *user stereotype*, in contrast, represents a certain kind of user who exhibits a set of specific characteristics. One approach of constructing user stereotypes is manual creation, based on e.g. age, sex, or any other feature or combination thereof. User stereotypes may also be identified by using clustering techniques to group similar users and identifying the key aspects of their similar features.

One of the primary advantages of utilizing user stereotypes is that before knowing a new user to the full extent, a system can make early assumptions about which type of user he/she is, based on the currently available personal information. Thus, qualified guesses can be made regarding which kind of action should be appropriate to satisfy a particular user [MOBASH2001] or strengthen a hypothesis in a medical context.

As introduced by Rich in [RICH1979], user stereotypes require two types of information. The system must know what properties capture a stereotype, and what events or behavior that implies a particular stereotype. If this information is highly dynamic and domain dependent, a clustering approach is preferable to static stereotypes, since it is able to automatically identify related categories and adapt to a changing population of users, their preferences and their characteristics.

By representing a solution to the problem of supplying a ‘typical’ kind of user with appropriate information and/or treatment, it is natural to see user stereotype cases as part of a Case-Based Reasoning process. When the information in a single user model is insufficient for deciding which items to select, the user stereotype case most closely resembling the user is consulted to make assumptions about the user’s expected behavior (Retrieve, Reuse). The case is revised when the user evaluates the recommended items, and Retained when the user stereotype cases are updated.

### 2.3 Clustering

Clustering is a type of multivariate statistical analysis also known as cluster analysis, or unsupervised classification analysis. Clustering is used to group items/persons/etc. into separate clusters based on their statistical behaviour. The main objective of clustering is to find similarities between samples, and then group similar samples together to assist in understanding relationships that might exist among them.

Cluster analysis is based on a mathematical formulation of a measure of similarity. There are a number of characteristics that distinguish different approaches to cluster analysis.

- Numerical, statistical, and conceptual clustering.
- Agglomerative vs. divisive.
- Overlapping vs. disjoint clusters.
- Incremental vs. non-incremental.
- Flat vs. hierarchical representations.

To measure similarity, a distance metric is used. There are a number of different distance metrics that are often used, and they are separated into *Distance Measurements Between Data Points,* and *Distance Measurements Between Clusters.* The most common of the first kind are Euclidian distance, Manhattan distance, Pearson correlation distance, and Spearman distance. The latter are divided into average linkage, single linkage, and complete linkage.

Three of the most common hierarchical methods are k-means, agglomerative hierarchical clustering, and Self-Organizing Maps (SOM).

A problem inherent in every clustering method is the problem of choosing the optimal number of clusters. Choosing too many clusters compromises generality, but choosing too few clusters may result in less distinct, less informative cluster groups. Two methods often used for determining the number of clusters are MDL (Minimum Description Length) and BIC (Bayes Information Criterion) [JAIN1988] [FUNG2001].

## 3 Motivation

This section is split into separate parts due to differing motivations in the two application domains – personalisation and psychophysiology. The motivation is different in the sense that user stereotypes and category-based classification are used to resolve different types of problems. The separation of application domains however helps to show the versatility of the proposed approach, and the way of dealing with these two domains are indeed identical enough to motivate a similar, more general problem-solving approach towards them.

### 3.1 Personalisation

In the case of web page filtering and personalisation, the need for category-based classification arose as a way of dealing with a very specific problem known as the *latency* problem. In short, this problem can be described as the problem of dealing with users who are new to a system/web site and for whom there is therefore insufficient information available to make decisions about proper actions based on the knowledge of that single user. By clustering and classifying data into categories, a user can quickly be classified as most similar to a specific user stereotype, and the knowledge about the stereotype can then be used to make assumptions about the expected behaviour of the user, due to the behaviour of similar users contained as probabilistic relationships within the user stereotype. It also included additional benefits, such as faster response times due to the offline nature of the clustering and data mining techniques.

### 3.2 Psychophysiology

The major difference between the motivation in the personalisation and the psychphysiological domain is the nature of the problem. In personalisation, the problem was clearly stated from the beginning, and most parameters were known or believed to be known. In the medical domain, and perhaps even more so in the psychophysiological domain, many parameters and their relations are largely unknown. Therefore, the later work is oriented more towards data *exploration*, in addiction to using what is already known. To find relationships between measurements variables, symptoms etc among patients with stress related diseases is actually motivation enough, since parts of the field are virtually unexplored. However, it is also essential to actually evaluate the assumptions made from data mining to make sure that they hold in reality.

The main motivation behind the research in the psychophysiological domain is to develop a methodology as well as a working prototype to advise physicians on what tests should be completed to classify a patient into one or more disease groups. Making tests are both time consuming and costly, and by limiting tests to only those that are strictly necessary, diagnosis costs can be reduced.

## 4 Related work

Introduced early on by Rich [RICH1979] and further developed by Rich in [RICH1989], user stereotypes have often been employed in the user modelling community, e.g. by Paliouras et.al. [PALIOURAS99] to model users in a dialog system, by Jameson [JAMESON1992], applying a psychological perspective, by Chin [CHIN89] who explore the advantages of user stereotypes compared to user models, and by Dailey [DAILEY96], who interestingly use stereotypes as a way of handling bias in statistical data.

The usage of stereotypes is also very common in information filtering on the web, as a way of classifying users. Ardissino and Sestero [ARDISSINO] uses stereotypes as a way of modelling user plans. Kuflik et.al. [KUFLIK03] provides a thorough comparison of information filtering systems based on user stereotypes versus filtering systems based on personal, individual information. Henze and Nejdl [HENZE03] utilize stereotypes in an online educational system to provide better learning.

CBR is commonly used in information filtering alongside filtering strategies such as collaborative and content-based filtering. It has been argued by Hayes et.al. [HAYES01] that under certain conditions, collaborative filtering and CBR can be seen as synonymous.

Computer-aided medical diagnose systems have been around since the 1970's, the first system being MYCIN, a system to diagnose blood infections [BUCHANAN84]. These early systems were generally completely rule-based, and although sometimes functioning fairly well, suffered from severe maintenance and rule inconsistency problems [SPANGLER92]. There are also ethical considerations that were and still are valid, as described by Spyropoulos in [SPYROPOULOS98].

Stereotypes have not been used in the domain of medical diagnosis. In fact, CBR is in itself a relatively new field of research within medical diagnose systems research. For examples of the usage of CBR in medical diagnosis, see [BRADBURN93] [SCHMIDT00] [GIERL98]. One of the earliest medical expert systems utilizing CBR was CASEY [KOTON88], that combines a CBR-approach with a model based expert system for diagnosing cardiac diseases (heart failure).

Although there have been many attempts at creating fully functional diagnose systems, experts are typically still a crucial part of the decision chain at later stages. The principal, emerging value of computers in medicine over the past several decades has first and foremost been one of organizing and communicating fussy, detailed information about patients, such as physician-orders and medical records (MURFF01).

There have been no reported attempts at creating a medical diagnosis system for the particular task of diagnosing patients with stress related diseases and/or symptoms. Two examples that deal with vaguely similar topics are Montani [MONTANI01], who in his Ph.D. thesis explores the use of decision support in diabetes care, and Marling and Whitehouse [MARLING01], who examine the possibilities of using CBR for prescribing drugs to patients suffering from Alzheimer's disease.

## 5 Contribution

The main planned contributions of the thesis are summarized below.

**General framework for category-based classification and user stereotypes**. A description and specification of the proposed approach to extend it to a general domain.

**Algorithms for classification and dialog-based category-based filtering**. Special purpose algorithms to handle specific problems related and specific to the addressed domains, while being kept as general as possible not to narrow it unnecessarily.

**Empirical evaluation of real data sets.** The presented methods, in particular category-based diagnosis and classification, will be used on available data sets to prove the concept. Due to the earlier described shift in interest and projects, the evaluation will be focused on the medical domain.

## 6. Thesis outline

In this section the outline for the forthcoming licentiate thesis is proposed.

1. Introduction
This section introduces the field of research, terminology and background.

2. Methodology
In this section, we cover background material and the general methodology for the user stereotype and category-based classification approach.

3. Related work
The work relevant to the thesis is referred. Similarities and differences between referred paper and the thesis are discussed.

**4. Paper A.** Mikael Sollenborn and Peter Funk: Category-Based Filtering and User Stereotype Cases to Reduce the Latency Problem in Recommender Systems. This paper has been published and presented at the ECCBR 2002 conference in Aberdeen, Scotland.

The paper describes a web page personalization approach referred to as *category-based filtering*. Its main characteristic is that selection of information is based on category ratings instead of item ratings, in contrast to other content-filtering strategies in general, and representationless collaborative filtering in particular. The selection of items is based partly on individual user models, and partly on collective user stereotypes cases. A *user model* represents the current knowledge about a user's reaction towards shown categories of items. A *user stereotype case*, in contrast, consists of collective information about a group of users.

The main contribution of this paper is to show how user stereotypes and CBR can be used in the context of web page filtering. It contains the first steps toward a general framework for category-based classification and user stereotypes.

**5. Paper B.** Technical report: User stereotypes for efficient classification. This report will act as a bridge between Paper A and Paper B, explaining the general features of user stereotypes, clustering, and classification. This is basically the SOTA report used as a technical report.

**6. Paper C.** Patient stereotypes for category-based **s**ymptom matching and prediction. This paper extends the stereotype and category-based classification to the medical domain, Building on the knowledge in Paper B, and parts of the methodology in Paper A, the paper attempts to prepare for answering the following question: which questions/measurement do I need to ask/perform to clarify to which patient stereotype a particular patient belongs?

**7. Paper D.** Using patient stereotypes for category-based **s**ymptom matching and prediction: an evaluation. Building on the methodology developed in Paper C, this paper contains an evaluation of using an automated symptom diagnose for stress patient treatment.

8. Conclusion and future work
This part summarizes the main contributions of the thesis, presents conclusions of the theoretical framework and experimental evaluation, and possible directions of the future research.

## 7. Time plan

This section contains the proposed time plan for the licentiate thesis project.

### 7.1 Completed work

- Written and published a paper on category-based filtering and user stereotype cases for reducing the latency problem in Recommender Systems. Published in full version in the Proceedings of the 6th European Conference on Case-Based Reasoning (ECCBR) 2002. Also published earlier on as a short paper and in different form, in the Proceedings of the 2nd International Conference on Adaptive Hypermedia and Adaptive Web Based Systems (AH) 2002.

- Co-written and published a paper on classification of complex measurements using CBR. Published in the Workshop Proceedings of the workshop on health sciences at the 5th International Conference on Case-Based Reasoning (ICCBR) 2003.

- Completed courses:
  Research methodology for computer science and engineering (5p); Science planning for Ph.D. students (5p); Artificial Intelligence Advanced course (5p); Multi agent systems (5p); Artificial Intelligence (5p).

- A prototype on category-based filtering and clustering of users on web sites.

### 7.2 Remaining work

| | | |
|---|---|---|
| October | 2003 | Licentiate proposal. Take a 5p course in psycho physiological medicine. |
| November | 2003 | Evaluating, using clustering, patterns in stress patient Journals. Take a 5p course in CBR. |
| December | 2003 | Paper C finished and submitted for review |
| January | 2004 | Theoretically extending the general framework |
| February | 2004 | Verification of algorithms and framework for category-based symptom classification |
| March | 2004 | Paper D finished and submitted for review |
| May | 2004 | Paper B finished |
| June | 2004 | Licentiate thesis draft ready for review |
| July | 2004 | Licentiate thesis finished and presented |

## 8. Future work

In this section, several future research directions in the area of user stereotypes and category-based classification will be discussed.

**Further exploration and evaluation of stress patient parameter relationships.**

Psychophysiology is a vast area, where, as said before, much is unknown. The above referred (Paper D), exploring symptom relations, is only a very small step towards understanding the psychophysiological motivations behind stress related diseases. By digging into the vast amount of patient data available, further interesting data relationships and facts will hopefully be explored and utilized.

**"Online" evaluation of stress measurement data.** With the coming wireless sensors for measuring stress patients, there will no longer be a need to isolate measurements into distinct sessions where measurements are performed by a medical expert. This will lead to a number of advancements, but also to interesting new problems. The constantly ongoing measurements will make it possible to examine the dynamic changes of bodily functions over time and during different types of situations. This will lead to enormous amounts of data, that quite possibly must be interpreted in new ways due to its dynamic nature.

**Building a complete stress patient diagnose aiding system.** Building on the sensor reading and analysis of sensor data continually developed by colleague Markus Nilsson, pattern analysis of measurement data will be merged with patient specific data and Case-Based Reasoning methodology, hopefully in the end producing an integrated, highly autonomous stress diagnose system.

## References

[RICH79] Rich, E.*: User Modeling via Stereotypes.* Cognitive Science 3 (1979) 329-354

[RICH89] Rich E. (1989): Stereotypes and User Modeling. In A. Kobsa and W. Wahlster, editors, UserModels in Dialog Systems, pages 31-51. Springer Verlag, Berlin

[KOLOD93] Kolodner, J.L. (1993). *Case-Based Reasoning*. San Francisco: Morgan Kaufmann Publishers.

[WATSON97] Watson, I. (1997). *Applying Case-Based Reasoning: Techniques for Enterprise Systems*, Morgan Kaufmann Publishers.

[MOBASH01] B. Mobasher, H. Dai, T. Luo, M. Nakagawa. Improving the Effectiveness of Collaborative Filtering on Anonymous Web Usage Data. In workshop *Intelligent Techniques for Web Personalization*, editors S. Anand, B. Mobasher, pp 53-60, IJCAI-2001, Seattle Washington.

[JAIN88] Jain, A.K., Dubes, R.C.: *Algorithms for Clustering Data*. Prentice Hall (1988)

[FUNG01] Fung,, Glenn. A Comprehensive Overview of Basic Clustering Algorithms, May 2001.

[HENZE03] Nicola Henze and Wolfgang Nejdl. Logically characterizing adaptive educational hypermedia systems. In Proc. of the AH'2003.

[ARDISSINO96] L. Ardissono and D. Sestero. Using dynamic user models in the recognition of the plans of the user. User Modeling and User-Adapted Interaction, 5(2):157--190, 1996.

[KUFLIK2003] Kuflik , T., Shapira, B., Shoval, P. " Stereotype-Based Vs. Personal-Based Filtering Rules in Information Filtering Systems". Journal of American Society of Information Systems Technology (JASIST) 2003

[PALIOURAS99] G. Paliouras, V. Karkaletsis, C. Papatheodorou, C.D. Spyropoulos, "Exploiting Learning Techniques for the Acquisition of User Stereotypes and Communities," Proceedings of the International Conference on User Modelling (UM '99).

[MURFF01] Murff HJ, Kannry J. Physician satisfaction with two order entry systems.
J Am Med Inform Assoc. 2001 Sep-Oct;8(5):499-509.

[JAMESON92] Jameson, A. Generalizing the double-stereotype approach: A psychological perspective. In UM92: Third International Workshop on User Modelling., 69-83., 1992.

[CHIN89] Chin, D.N., KNOME: Modelling what the user knows in UC. In Kobsa, A. and Wahlster, W. eds. User Models in Dialog Systems. Berlin: Springer-Verlag, 133-162, 1989.

[DAILEY96] Matthew N. Dailey, Gregory S. Miller, and James C. Lester Exploiting Stereotypes to Eliminate Strategic Bias, Proceedings of the 5th International Conference on User Modeling, Kailuna-Kona, Hawaii, USA, 2-5 Jan. 1996

[KOTON88] Koton, P., Using Experience in Learning and Problem Solving. MIT Press, 1988.

[HAYES01] Hayes, C., Cunningham, P., Smyth, B.: A Case-Based Reasoning View of Automated Collaborative Filtering. In Proceedings of 4th International Conference on Case-Based Reasoning, ICCBR2001 (2001) 243-248

[BUCHANAN84] Buchanan, B. G., and E. H. Shortliffe. Rule-Based Expert Systems: The MYCIN Experiments oJ the Stanford Heuristic Programming Project. Reading, MA: Addison-Wesley, 1984.

[MARLING01] Marling, C., and Whitehouse, P., Case-based Reasoning in the care of Alzheimer's disease patients. In Case-Based Reasoning and Development, pp. 702-715. 4th International Conference on Case-Based Reasoning, (2001).

[SPANGLER92] Spangler, William E., May, Jerrold H. Success and failure in cooperative expert systems development: a tale of two projects. Source Journal of Systems and Software archive. Volume 19 , Issue 2 (October 1992) Special issue on expert systems that failed. Pages: 131 – 140., 1992

[SCHMIDT2000] Schmidt, R., Gierl L.: Case-based Reasoning for Medical Knowledge-based Systems. Proceedings of Medical Infobahn for Europe (2000)

[SPYROPOULOS98] Spyropoulos B., Papagounos G.: Ethical aspects of the employment of Expert Systems in Medicine. Proceedings of the 4th International Conference on Ethical Issues of Information Technology, p. 701-711. (1998)

[GIERL98] Gierl L., et al: CBR in Medicine, Case-Based Reasoning Technology. From Foundations to Applications, pp. 273-297 (1998)

[BRADBURN93] Bradburn C., Zeleznikow J.: The application of case-based reasoning to the tasks of health care planning, Proceedings of European Workshop on CBR, pp.365-378 (1993)

[MONTANI01] Montani S.: Knowledge Management and Decision Support in Diabetes Care through Multi Modal Reasoning. PhD thesis, University of Pavia (2001)