DISS. ETH NO. 20600

INFORMATION IN ORDERINGS (LEARNING TO ORDER)

A dissertation submitted to
ETH ZURICH

for the degree of
Doctor of Sciences

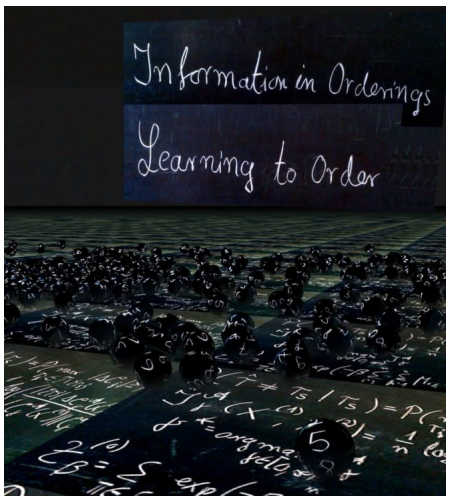
presented by
LUDWIG MAXIMILIAN BUSSE
Master of Science ETH in Computer Science
born May 30th, 1984
citizen of Germany

accepted on the recommendation of
Prof. Dr. Joachim M. Buhmann, examiner
Prof. Dr. Peter Widmayer , co-examiner

2012

INFORMATION IN ORDERINGS

LUDWIG MAXIMILIAN BUSSE



Learning to Order

August 2012

ABSTRACT

This thesis is concerned with *order information*. Many data come in the form of a *partial* or *total order* rather than as points in an Euclidean space. Rankings, for example, are used to indicate the relative importance of websites, universities, or sports teams. Humans also like preference rankings to state their tastes on e.g. music or clothes. And computers spend a massive amount of their time sorting data, that is, establishing order.

This work begins with studying statistical models for *analyzing rank data*. A model family for cluster analysis of heterogeneous rank- and pairwise comparison data is developed.

Cluster analysis of ranking data, which occurs in consumer questionnaires, voting forms or other inquiries of preferences, attempts to identify typical groups of rank choices. Empirically measured rankings are often incomplete, i.e. different numbers of filled rank positions cause heterogeneity in the data. We propose a mixture approach for clustering of heterogeneous rank data. Rankings of different lengths can be described and compared by means of a single probabilistic model. Thereby, an unsupervised clustering approach for rank data is presented that is capable of performing an integrated analysis on heterogeneous data with different patterns of missings.

Next, paired comparison models are defined for partial orders, where respondents decide between just 2 items at a time. With many items (e.g. products), we must expect the stated preference data to be incomplete: Many people rank a different subset of items. Various patterns of incomplete data can be accommodated, whether built into the design of an experiment or occurring by chance.

A method for multicriteria scaling is developed that makes it possible to reconstruct the latent utility weights that a society attributes to different decision options based on stated preference data. Our model can reconcile inconsistent choice (intransitivities), and thus in practice can lead to insights into decision making processes.

In total, a nested family of models with increasing complexity is presented in order to find structure in rank data. It is instantiated for *preference modeling* (preference elicitation, clustering, multicriteria decision making).

This thesis aims at sheding light on the notion of *in-form-ation*, here as related to the *ordering* meaning. An attempt to a context-sensitive measure of order information is established. Sorting algorithms (which produce an output ordering) are analyzed with a new methodology regarding as to how much order information they can generate in noisy settings, giving new insights into the informativeness of standard well-known sorting algorithms. This method, culminating in the so-called *Generalization Capability* of algorithms, opens a new dimension for the analysis of algorithms beyond time- and space considerations by answering the question: How many task-related bits can an algorithm extract from the input data?

A practical result includes the derivation of a new sorting algorithm that behaves optimal under uncertainties in the input data or in primitive algorithmic operations.

ZUSAMMENFASSUNG

Diese Arbeit beschäftigt sich mit *Ordnungs-Information*. Viele Daten treten nicht als Punkte in einem Euklidischen Raum, sondern in Form einer *partiellen* oder *totalen Ordnung* auf. Beispielsweise werden *Rankings* verwendet, um die relative Wichtigkeit von Webseiten, Universitäten oder Sportteams auszumachen. Menschen mögen sog. Präferenzrankings, um ihre Vorlieben bezüglich z.B. Musik oder Bekleidung auszudrücken. Und Computer verbringen einen Grossteil ihrer Zeit damit, Daten zu sortieren, also Ordnungen herzustellen!

Diese Arbeit beginnt mit dem Studium von statistischen Modellen zur Analyse von Rankdaten. Es wird eine Modellfamilie für die Clusteranalyse von heterogenen Rang- und Paarvergleichsdaten entwickelt.

Eine Clusteranalyse von Rankdaten (wie sie in Konsumentenfragebögen, Wahlscheinen oder anderen Präferenzerhebungsmethoden vorkommen) zielt darauf ab, typische Gruppen von Rangentscheidungen zu identifizieren. Empirisch erhobene Rankings sind oftmals unvollständig, d.h. eine unterschiedliche Anzahl von besetzten Rängen verursacht Heterogenität in den Daten. In dieser Arbeit wird ein Mixturvorgehen vorgeschlagen, um heterogene Rankdaten zu clustern. Rankings verschiedener Längen können mit einem einzigen probabilistischen Modell beschrieben und verglichen werden. Es wird also ein unüberwachtes Lernverfahren präsentiert, welches eine integrierte Analyse von heterogenen Daten mit unterschiedlichen Mustern von fehlenden Daten erlaubt.

Als nächstes wird ein Modell für paarweise Vergleiche definiert (über partielle Ordnungen), wobei Ranker nur zwischen jeweils 2 Optionen gleichzeitig entscheiden müssen. Sobald viele Optionen zur Auswahl stehen, werden die zum Ausdruck gebrachten Präferenzen häufig unvollständig sein: Viele Antwortgeber ranken nur eine bestimmte Anzahl von Alternativen, und diese unterscheiden sich zwischen den Rankern auch noch. Unser Modell kann mit ganz unterschiedlichen Mustern von Unvollständigkeiten umgehen, egal ob diese im Experimentdesign be-

absichtigt waren oder zufällig auftreten.

Weiter wird eine Methode zur Multikriterien-Skalierung entwickelt. Sie ermöglicht es, die nicht offenliegenden Nützlichkeitswertigkeiten zu ermitteln, welche eine Gesellschaft verschiedenen Entscheidungsoptionen beimisst (basierend auf geäusserten Präferenzaussagen). Dieses unser Modell kann auch mit inkonsistenten Auswahlen umgehen (z.B. Intransitivitäten), und somit zu Erkenntnissen in schwierigen Entscheidungssituationen führen.

Insgesamt wird eine Familie von Modellen mit zunehmender Komplexität präsentiert, um Struktur in Rangdaten zu finden. Diese Modelle werden anhand des Beispiels "Präferenzmodellierung" (Präferenzabfrage, Gruppierung, Entscheidungsfindung unter mehreren Kriterien) zur Anwendung gebracht.

Diese These hat als Ziel, eine Notation von "In-form-ation" zu entwerfen (und zwar am Beispiel der "Ordnungs"-Bedeutung). Eine Möglichkeit für ein kontext-sensitives Informationsmass für Ordnungsinformation wird gegeben. Sortieralgorithmen (die eine Ordnung als Ausgabe produzieren) werden mit der neuen Technik analysiert in Bezug darauf, wieviel Ordnungsinformation sie in unsicheren Situationen etablieren können. Dies führt zu neuen Erkenntnissen über die "Informativität" von wohlbekannten Standard-Sortieralgorithmen. Das Vorgehen findet seinen Höhepunkt in der sog. *Generalisierungs-Kapazität* von Algorithmen, die eine neue Analysedimension neben Zeit- und Speicherplatzanforderungen eröffnet, indem die Frage beantwortet wird: Wieviel Aufgaben-bezogene Bits kann ein Algorithmus aus den Eingabedaten extrahieren?

Eine praktische Folge ist die Herleitung eines neuen Sortieralgorithmus, der sich unter den Unsicherheiten in der Eingabe oder in seinen primitiven atomaren Operationen optimal verhält.