

STATISTISCHES CONSULTING

MASTER STUDIENGANG STATISTIK

INSTITUT FÜR STATISTIK

LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Statistisches Consulting für die Interhyp AG Marketing im Internet

Daniel Fuckner d.fuckner@gmx.de
Markus Vogler markus@vogler-lindau.de

Projektpartner:
Interhyp AG

Betreuer:
Dr. Fabian Scheipl

München, 11.11.2011

Abstract:

Background: In patients with

Inhaltsverzeichnis

1	Einleitung	1
2	Datenlage	2
3	Zeitdiskretes Proportional-Hazards-Modell von Cox	2
4	Sequential Pattern Mining	2
4.1	Überblick	2
4.2	Problemstellung	2
4.3	Auswahl geeigneter Algorithmen	3
4.3.1	Apriori-Based	3
4.4	Pattern-Growth	4

1 Einleitung

Die Interhyp AG ist Vermittler für private Baufinanzierungen. Das heißt, sie wählt aus einem Angebot von verschiedenen Darlehensgebern die optimale Finanzierungsstruktur für einen Kunden aus. Das Unternehmen wurde 1999 basierend auf der Idee, die Baufinanzierungsbranche zu revolutionieren, von den ehemaligen Goldman-Sachs-Bankern Robert Haselsteiner und Marcus Wolsdorf gegründet. Sechs Jahre später eröffnete die Interhyp AG erste Niederlassungen und konnte gleichzeitig den erfolgreichsten deutschen Börsengang des Jahres verzeichnen. Nach weiteren drei Jahren erfolgte die Übernahme durch ING DIRECT, der weltweit größten und erfolgreichsten Direktbanken-Gruppe. Heute ist die Interhyp AG der größte Vermittler für private Baufinanzierungen in Deutschland, wurde acht mal in Folge als "Bester Baufinanzierer" (Zeitschrift *€*, Ausgabe 08/2013) ausgezeichnet und verfügt über mehr als 60 Beratungsstandorte mit über 1.000 Mitarbeitern.

Das primäre Ziel des Marketing der Interhyp AG ist die Kundenakquise. Da etwa 80% aller Kundenanträge online abgeschickt werden, liegt der Fokus der Marketing-Abteilung auf dem Online-Marketing, das über verschiedene Kanäle verfügt. Beispiele sind die Kooperationen (z.B. mit Immobilienscout24), Suchmaschinen (bezahlte Anzeigen und unbezahlte Ergebnisse), Affiliate Marketing (Netzwerk kleinerer Partnerseiten), Display Advertising (diverse Bannerschaltungen), Newsletter und Social Media (vorrangig Facebook und gutefrage.net). Durch Online-Tracking können die Werbekontakte eines potentiellen Kunden mit der Interhyp AG zusammengefasst werden. So entsteht ein Customer Journey (siehe Abbildung ??), dass zum Abbruch oder im Idealfall zum Ausfüllen eines Onlineantrages führt. An dieser Stelle kommt die Refined Labs GmbH ins Spiel.

Customer Journey Grafik aus Folien

Die Refined Labs GmbH ist auf dem Gebiet des Online-Marketing spezialisiert und führender Anbieter für Performance-Marketing-Software. Zum Kundenportfolio zählt unter anderem auch die Interhyp AG, das heißt das Online-Tracking wird von der Refined Labs GmbH durchgeführt und verwaltet. Ein Customer Journey beginnt mit dem ersten Online-Werbekontakt eines potentiellen Kunden mit der Interhyp AG und der damit einhergehenden Erstellung eines Cookies. So können alle weiteren Werbekontakte dem potentiellen Kunden eindeutig zugewiesen werden. Das Tracking endet sobald der potentielle Kunde einen Onlineantrag versendet und damit zum Kunden wird. Wird innerhalb von 90 Tagen kein Onlineantrag versendet, so wird das Cookie automatisch gelöscht und man spricht von einem Abbruch.

Die Interhyp AG ist daran interessiert, ob ein Abbruch von Customer Journeys verhindert werden kann. Hier dann die Zielsetzungen einfügen, die wir gelöst haben.....

Überblick über alle folgenden Kapitel.

Quellen: Projektausschreibung und Folien von Frau Gries; wie angeben?

2 Datenlage

3 Zeitdiskretes Proportional-Hazards-Modell von Cox

Aufgrund der in Kapitel 2 beschriebenen Datenlage erscheint die Anwendung eines Modells aus dem Feld der Lebensdaueranalyse intuitiv. Für solch ein Regressionsmodell müssen die Daten in der Form $(t_i, \delta_i, x_i(t))$, $i = 1, \dots, n$ vorliegen.

4 Sequential Pattern Mining

4.1 Überblick

Überarbeiten!!!

Sequential pattern mining entdeckt häufige *subsequences* (dt. Teilfolgen) in Datenbanken. Sogenannte *sequence databases* bestehen aus Transaktionen, die jeweils *items* enthalten, welche der Zeit nach geordnet sind. Die Daten lassen sich also mit dem Schema [Transaction/ID, <Ordered Sequence Items>] darstellen.

Ein Anwendungsfeld ist die Warenkorbanalyse. Angenommen es wird das Kaufverhalten in einem Supermarkt einen Monat lang beobachtet, dann könnte [Kunde 1, <(Brot, Milch), (Brot, Milch, Tee), (Zucker), (Milch, Salz)>]; [Kunde 2, <(Brot), (Milch, Tee)>] eine Beispiel-Datenbank sein. Kunde 1 war vier mal im beobachteten Monat im Supermarkt einkaufen, wobei Kunde 2 nur zweimal einkaufen war. Der Kunde kann nur eins oder auch mehrere *items* pro Besuch einkaufen. Im Falle von mehreren *items* spricht man von *itemsets*.

Web usage mining ist das am weitesten verbreitete Anwendungsfeld von *sequential pattern mining* in der Literatur ([8, 16, 6]). Unter der Annahme, dass ein Internetnutzer nur eine Webseite an einem Zeitpunkt aufrufen kann, besteht die Folge von geordneten *items* nur aus einzelnen *items* und nicht aus *itemsets*. Ist also eine Menge von *items* $I = \{a, b, c, d, e\}$ gegeben, die beispielsweise verschiedene Webseiten repräsentieren, so könnte eine Datenbank mit zwei Nutzern folgendermaßen aussehen: [Nutzer 1, <abedcab>]; [Nutzer 2, <edcaa>] ([9, 3:1-3:2]).

4.2 Problemstellung

Das Problem wurde erstmals im Jahre 1995 in der Arbeit *Mining Sequential Patterns* [1] von Agrawal & Srikant vorgestellt. Es ist eine Datenbank D von Transaktionen gegeben, wobei jede Transaktion eine ID, den Zeitpunkt der Transaktion und die dazugehörigen *items* enthält. Für jede ID können mehrere Transaktionen existieren allerdings nie zwei oder mehr Transaktionen mit dem selben Zeitpunkt.

Ein itemset $i = (i_1 i_2 \dots i_m)$ ist eine nichtleere Menge von *items* i_j und eine Sequenz $s = \langle s_1 s_2 \dots s_n \rangle$ eine geordnete Liste von itemsets s_j . Ohne Beschränkung der Allgemeinheit wird angenommen, dass die Menge der *items* auf eine Menge von fortlaufenden ganzen Zahlen abgebildet wird.

Eine Sequenz $\langle a_1 a_2 \dots a_n \rangle$ heißt Subsequenz einer anderen Sequenz $\langle b_1 b_2 \dots b_m \rangle$ wenn ganze Zahlen $i_1 < i_2 < \dots < i_n$ existieren, so dass $a_1 \subseteq b_{i_1}, a_2 \subseteq b_{i_2}, \dots, a_n \subseteq b_{i_n}$ gilt. Beispielsweise ist $\langle (1)(58)(27) \rangle$ Subsequenz von $\langle (15)(248)(358)(27) \rangle$, da $(1) \subseteq (15), (58) \subseteq (358)$ und $(27) \subseteq (27)$ aber $\langle (27) \rangle$ keine Subsequenz von $\langle (2)(7) \rangle$ und umgekehrt. Eine Sequenz heißt maximal in einer Menge von Sequenzen, wenn sie von keiner Sequenz in dieser Menge Subsequenz ist.

Die Transaktionen einer ID können als Sequenz verstanden werden. Jede Transaktion entspricht dabei einem itemset und die nach den Zeitpunkten T_1, T_2, \dots, T_n geordneten Transaktionen entsprechen einer Sequenz $\langle itemset(T_1) itemset(T_2) \dots itemset(T_n) \rangle$. Der support einer Sequenz s ist der Anteil der IDs, die s unterstützen, das heißt deren Sequenz Subsequenz von s sind. Eine Sequenz mit k items wird auch k -Sequenz genannt. Das Problem des sequential pattern mining besteht darin, in einer Datenbank D unter allen existierenden Sequenzen die maximalen Sequenzen zu finden, deren support größer oder gleich einem festgelegten minimalen support $min_sup \ \xi$ ist.

4.3 Auswahl geeigneter Algorithmen

In den letzten zwei Jahrzehnten wurden im Forschungsfeld des sequential pattern mining eine Vielzahl von Algorithmen entwickelt ([14, 17, 1, 15, 10, 2, 7, 12, 11, 4, 3, 18, 5]; **die die an anderer stelle zitiert werden, hier später löschen**). Im Folgenden werden verschiedene Arten von Algorithmen sowie Merkmale anhand derer deren Qualität bewertet werden kann vorgestellt. Anschließend wird ausgearbeitet, welche dieser Algorithmen für die vorliegenden Daten am besten geeignet sind. Die Algorithmen lassen sich in die Kategorien *apriori-based*, *pattern-growth* und *early-pruning* einteilen, wobei es auch hybride Algorithmen gibt.

4.3.1 Apriori-Based

Apriori-basierte Algorithmen waren Mitte der neunziger Jahre die ersten auf dem Feld des sequential pattern mining und hängen vor allem von der apriori Eigenschaft ab. Diese besagt, dass alle nichtleeren Teilmengen eines häufigen itemsets ebenfalls häufig sein müssen. Sie können als monoton fallend verstanden werden da für jede Sequenz, die den minimalen support $min_sup \ \xi$ nicht erfüllt auch all deren Übermengen diesen nicht erfüllen. Das bedeutet, es gilt Gleichung (1).

$$X \subseteq Y \Rightarrow support(Y) \leq support(X) \text{ für } X, Y \subseteq J, J : \text{Menge aller items in } D \quad (1)$$

Damit sind für eine nicht häufige Sequenz auch alle Übermengen nicht häufig und für eine häufige Sequenz alle Untermengen ebenfalls häufig. Apriori-Algorithmen sind durch drei Hauptmerkmale gekennzeichnet. Ein Nachteil ist die *generate-and-test*-Eigenschaft. Sie bringt mit sich, dass ausgiebige join-Operatoren verwendet werden, die die Sequenzen in jedem Schritt um nur ein item vergrößern und gegen den minimalen support testen. Dadurch entsteht eine enorme Anzahl von Kandidat-Sequenzen, die bereits zu Beginn des

Algorithmus viel Speicher aufbrauchen. Sind beispielsweise n häufige 1-Sequenzen und $\min_{sup} = 1$ gegeben. Dann werden $n^2 + \binom{n}{2}$ Kandidat-2-Sequenzen und $\binom{n}{m}$ Kandidat- m -Sequenzen erzeugt.

4.4 Pattern-Growth

Anfang der 2000er Jahre entstand die *pattern growth*-Methode.

Literatur

- [1] R. Agrawal & R. Srikant. „Mining sequential patterns“. In: *Proceedings of the 11th Conference on Data Engineering (ICDE'95)* (1995), S. 3–14.
- [2] J. Ayres et al. „Sequential pattern mining using a bitmap representation“. In: *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2002), S. 429–435.
- [3] D.-Y. Chiu, Y.-H. Wu & A. L. P. Chen. „An efficient algorithm for mining frequent sequences by a new strategy without support counting“. In: *Proceedings of the 20th International Conference on Data Engineering* (2004), S. 375–386.
- [4] M. El-Sayed, C. Ruiz & E. A. Rundensteiner. „FS-Miner: Efficient and incremental mining of frequent sequence patterns in web logs“. In: *Proceedings of the 6th Annual ACM International Workshop on Web Information and Data Management* (2004), S. 128–135.
- [5] C. I. Ezeife, Y. Lu & Y. Liu. „PLWAP sequential mining: Open source code“. In: *Proceedings of the 1st International Workshop on Open Source Data Mining: Frequent Pattern Mining Implementation* (2005), S. 26–35.
- [6] B. Goethals. „Frequent set mining“. In: *The Data Mining and Knowledge Discovery Handbook* (2005), S. 377–397.
- [7] J. Han et al. „Freespan: Frequent pattern projected sequential pattern mining“. In: *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2000), S. 355–359.
- [8] Y. Lu & C. I. Ezeife. „Position coded pre-order linked WAP-tree for web log sequential pattern minings“. In: *Proceedings of the 7th Pacific-Asia Conference on Knowledge Discovery and Data Mining* (2003), S. 337–349.
- [9] Nizar R. Mabroukeh & C. I. Ezeife. „A Taxonomy of Sequential Pattern Mining Algorithms“. In: *ACM Computing Surveys* 43.1 (2010), 3:1–3:41.
- [10] F. Masseglia, O. Poncelet & R. Cicchetti. „An efficient algorithm for web usage mining“. In: *Network Inform. Syst. J.* 2 (1999), S. 571–603.
- [11] J. Pei, J. Hani, B. Mortazavi-Asl & H. Zhu. „Mining access patterns efficiently from web logs“. In: *Knowledge Discovery and Data Mining. Current Issues and New Applications. Lecture Notes Computer Science* 1805 (2000), S. 396–407.
- [12] J. Pei, J. Hani, B. Mortazavi-Asl & H. PINTO. „PrefixSpan: Mining sequential patterns efficiently by prefix-projected pattern growth“. In: *Proceedings of the International Conference on Data Engineering* (2001), S. 215–224.
- [13] R. Rymon. „Search through systematic set enumeration“. In: *Proceedings of the 3rd International Conference on Principles of Knowledge Representation and Reasoning* (1992), S. 539–550.

- [14] S. Song, H. Hu & S. Jin. „HVSM: A new sequential pattern mining algorithm using bitmap representation“. In: *Advanced Data Mining and Applications* 3584 (2005), S. 455–463.
- [15] R. Srikant & R. Agrawal. „Mining sequential patterns: Generalizations and performance improvements“. In: *Proceedings of the 5th International Conference on Extending Database Technology: Advances in Database Technology* 1057 (1996), S. 3–17.
- [16] J. Wang & J. Han. „BIDE: Efficient mining of frequent closed sequences“. In: *Proceedings of the 20th International Conference on Data Engineering* (2004), S. 79–90.
- [17] Z. Yang, Y. Wang & M. Kitsuregawa. „LAPIN: Effective sequential pattern mining algorithms by last position induction for dense databases“. In: *Advances in Databases: Concepts, Systems and Applications* 4443 (2007), S. 1020–1023.
- [18] M. J. Zaki. „SPADE: An efficient algorithm for mining frequent sequences“. In: *Mach. Learn.* 42 (2001), S. 31–60.