"Maximizing the Spread of Influence in a social network"

Zusammenfassung des Vortrags gehalten im Rahmen des Datamining Seminars 06
von Jan Koopmann

Einleitung:

Bei der Konzeptierung von Marketingkampagnen möchte man mit möglichst wenig Einsatz in Form von Geld möglichst viele Kunden gewinnen.
Viral Marketing ist eine mögliche Lösung für dieses Problem. Man arbeitet auf Basis von Empfehlungen. Das heißt es wird initial eine Gruppe von Personen ausgewählt, die man mit dem Produkt vertraut macht. Diese, die im optimalen Fall begeistert von diesem Produkt vielleicht ein neuer Kosmetikartikel sind, empfehlen es an Freunde weiter, die es dann eventuell ihrem Bekanntenkreis infizieren. Die Begeisterung verbreitet sich epidemieartig über das soziale Netzwerk weiter. Solche Methoden, können zum Beispiel bei der Verbreitung eines neuen Handys unter Jugendlichen oder bei politischen Bewegungen zum Tragen kommen. Diese Trends können aber auch wieder abebben. Der folgende Text befasst sich mit der Auswahl einer guten Startergruppe auswählt und mit verschiedenen Modellen, die ein soziales Netzwerk simulieren können.

1.Basismodelle

Ein soziales Netzwerk ist dabei ein gerichteter Graph G bei dem die Individuen durch Knoten dargestellt werden und die Beziehungen durch Kanten. Die Knoten sind dabei entweder aktiv oder inaktiv. Die Tendenz aktiviert zu werden nimmt dabei zu je mehr Nachbarn aktiv sind. Ein Einmal aktivierter Knoten bleibt aktiv.
Die initiale Menge an aktivierte Knoten wird mit A_0 bezeichnet.

Linear Treshold Modell:

Die Idee hinter dem Linear Treshold Modell ist die Annahme, dass sobald genügend Nachbarn eines Knotens v von dem Produkt überzeugt sind auch v aktiviert wird.
Jeder Knoten wählt zufällig einen Schwellwert $\Theta_v \in [0,1]$
Jede Kante ist mit einem Gewicht bv,w ausgestattet. Das angibt wie stark der Einfluss eines Nachbarns w auf den Knoten v ist, Die Summe der Gewichte der eingehenden Knoten ist dabei kleiner als eins. Ein Knoten v wird in Schritt t aktiv, wenn die Summe der Gewichte der in Schritt t-1 aktiven Nachbarn den Schwellwert erreicht oder überschreitet.

Independent Cascade Modell:

Beim Independent Cascade-Modell gibt es für jedes benachbarte Knotenpaar (v,w) eine Wahrscheinlichkeit p_vw, das v den Nachbarknoten w in t aktiviert, sobald er in t-l aktiviert wird. Ein Knoten hat für jeden Nachbarn nur einen Aktivierungsversuch. Auch beim Independent-Cascade Modell werden diskrete Schritte betrachtet.
Als Maß für den Einfluss einer Knotenmenge A dient die Funktion σ(A). Sie gibt an, wie viele Knoten am Ende des Prozesses aktiv sind. Das in *Influence Maximization Problem* ist die Auswahl einer k-elementige Menge an Knoten zu finden für die σ(A) maximiert wird.

2.Komplexität und Approximierbarkeit:

Bei beiden Modellen ist das finden einer optimalen Lösung NP-schwer. Allerdings ist nach Nemhauser, Wolsey und Fischer eine
Approximation per *Greedy Hill-Climbing* mit einem Faktor von 1-1/e≈63% möglich, sollte es sich bei σ() um eine nicht negative, submodulare monotone Funktion handeln
Sei f eine Funtion, die Teilmengen einer Grundmenge auf eine reele Zahl abbildet.
Dann ist f submodular, wenn die die folgende *Diminishing Returns*-Eigenschaft erfüllt ist:

$f(S\cup\{v\})-f(S)\geq f(T\cup\{v\})-f(T)\quad \forall v \text{ und } \forall S\subseteq T$

$f \text{ ist monoton, wenn. } f(S\cup\{v\})\geq f(S)$

Erweiterung der Einflussformel:

Man kann zusätzlich eine unterschiedliche Bewertung w_v der Knoten angeben. Vielleicht sind manche Personen profitabler und kaufen mehr von dem Produkt.
σ(A) ergibt sich dann aus den Beziehungen:

$\sigma_w(A)=\sum_{v\in B} w_v;\ B:=\text{Am Ende aktive Knoten}$

$w_v=1\ \forall v \Rightarrow \text{Basisfall}$

$\sigma_w()$ dann submodular, wenn auch $\sigma()$ submodular!

$\sigma_w()$ ist genau dann submodular, wenn auch σ() submodular , daher ist diese Erweiterung für die weitere Beweisführung nicht relevant.
Der Beweis, das sowohl das Independent Cascade Modell, als auch das Linear Treshold-Modell per Greedy Hill-Climbing approximierbar sind, wird im folgenden geführt.

Beweis der Submodurität von σ() des Independent Cascade-Modells:

Jeder Knoten v hat beim Independent Cascade-Modell eine einmalige Chance einen bestimmten Nachbarknoten w mit der Wahrscheinlichkeit $p_{v,w}$ zu aktivieren. Es wird für jede Kante im vorraus ein Münzwurf ausgeführt um zu bestimmen, ob ein Aktivierungsereignis eintritt. Aus diesen Ergebnissen wird ein Graph konstruiert, wobei Kanten mit erfolgter Aktivierung als *live-edges* bezeichnet werden. Kanten mit nichterfolgter Aktivierung als *blocked-edges*. Ein Knoten x wird genau dann aktiviert, wenn es einen Pfad von einem Knoten aus A_0 zu x gibt, der nur aus *live-edges* besteht.
Seien X das Ergebniss aller Münzwürfe, dann sei R(v,X) die Menge aller Knoten, die von v aus auf einem Pfad von live-edges erreicht werden können und $\sigma_X(A):=U_{v\in A}R(v,X)$

Behauptung: $\sigma_X()$ ist submodular für alle festen X.
und somit die Submodularität von σ_X
S,T seien Knotenmengen mit S subseteq T.
$\sigma_X(S\cup\{v\})-\sigma_X(S)$ ist die Anzahl der Elemente in $R(v,X)$, die nicht bereits in $U_{u\in S}R(u,X)$ enthalten sind.
Diese ist mindestens so groß wie die Anzahl der Elemente $\in R(v,X)$, die nicht in der größerer Menge $U_{u\in T}R(u,X)$ enthalten sind, Startknoten die in T nicht aber in S enthalten sind, durch S aktiviert werden könnten.

Daraus folgt $\sigma_X(S\cup\{v\})-\sigma_X(S)\geq\sigma_X(T\cup\{v\})-\sigma_X(T)$

$\sigma(A)=\sum_X P(X)*\sigma_X(A)$
σ ist submodular, da eine nicht negative lineare Kombination von submodularen Funktionen wieder submodular ist

Beweis der Approximierbarkeit des Linear Treshold-Modells:

Man definiert auch hier eine andere Sichtweise auf den Prozess um dann die Äquivalenz der Verteilung aktiver Knoten zu zeigen. Danach erfolgt der Beweis der Submodularität wie beim Independent Case-Modell.
Es wird angenommen, dass ein Knoten maximal eine eingehende Kante als live-edge auswählt. Die Wahrscheinlichkeit der Aktivierung durch w in Schritt t ist dabei b_{vw} und die

Wahrscheinlichkeit, dass keine Aktivierung erfolgt $1-\sum_{w} b_{v,w}$

Der Beweis wird zuerst für einen gewichteten azyklischen Graph geführt. Dabei erhält man aus der topologischen Sortierung die Reihenfolge der Knoten in ihrer Aktivierung.
Im Linear Treshold Modell erhält man als Wahrscheinlichkeit für die Aktivierung von v_i, wenn seine Nachbarn S_i aktiv sind $\sum_{w \in S_i} b_{v,w}$

Dies entspricht der Wahrscheinlichkeit, dass die von v_i gewählte live-edge in S_i liegt.

Linear Treshold: Äquivalenz im Allgemeinen:
Im allgemeinen Fall erfolgt eine Induktion über die Iterationen des Prozesses . A_t ist dabei die Menge der aktiven Knoten am Ende von Schritt t. Der Wahrscheinlichkeit, dass neuaktivierten Knoten den Grenzwert überschreiten helfen, wenn dieser vorher noch nicht erreicht wurde entspricht dabei:

$$\frac{\sum_{u \in A_t \setminus A_{t-1}} b_{v,u}}{1-\sum_{u \in A_{t-1}} b_{v,u}}$$

Von der live edge-Sicht aus werden schrittweise per Breitensuche von A aus erreichbare Knoten $A_1,A_2,...,A_t$ bestimmt.Für einen Knoten v der in Schritt t noch nicht als erreichbar gekennzeichnet wurde ist die Wahrscheinlichkeit dass v als

erreichbar in t+1 gekennzeichnet wird ebenfalls:

Beide Prozesse produzieren also dieselbe Verteilung aktiver Knoten.

Experimenteller Vergleich:
Zum Testen des Modelle, wurde ein Graph aus Physik Publikationen erstellt. Die Knoten stellen jedes Paper bei dem zwei Autoren haben, wird eine Kante zwischen deren Knoten eingefügt. Paralle Kanten werden dabei als stärkere soziale Verbindungen gewertete.
Verglichen wurde der Einfluss von Methoden zur Auswahl der Initialmenge. Neben dem Greedy-Hill-Climbing-Algorithmus wurde eine Auswahl nach Anzahl abgehender Kanten ("high degree"), Zentralität der Knoten("central")und als Qualitätsmaß eine zufällige Auswahl.

$\frac{\sum_{u \in A_t \setminus A_{t-1}} b_{v,u}}{1-\sum_{u \in A_{t-1}} b_{v,u}}$ den Coautorenschaft in dabei Autoren dar. Für zusammengearbeitet

Die tatsächliche Qualität ist aufgrund der der Größe des Graphen und der NP-schwere des Problems nicht überprüfbar.

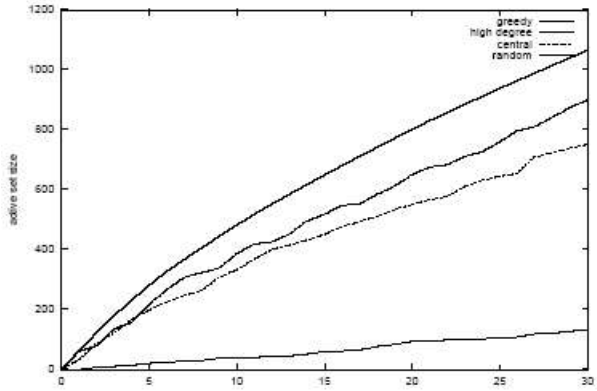

Abbildung 1:Linear Treshold: Abbildung entnommen aus [1]

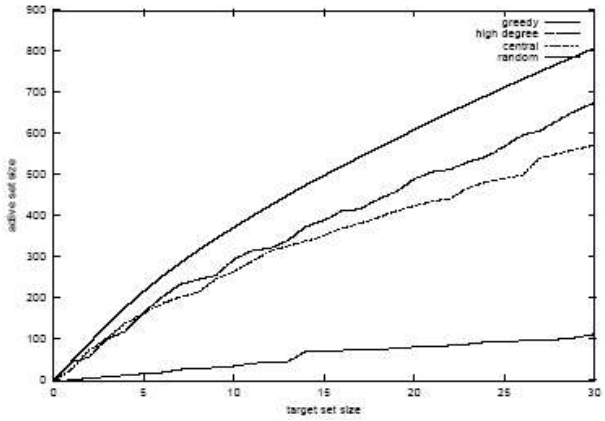

Abbildung 2:Weighted Cascade: Wahrscheinlichkeit der Aktivierung eines Knotens= 1/Anzahl der eingehenden Kanten. Abbildung entnommen aus [1]

Sowohl beim Linear Treshold, als auch beim Weighted Cascade, schneidet der Greedy-Algorithmus besser ab, wobei er beim Weighted Cascade-Modell, etwa 25% weniger effizient arbeitet. Eine Erklärung für das schlechtere Abschneiden der beiden anderen Modell ist die Klumpenbildung. Vielleicht gibt es Verdichtungen im Netz, in denen die Ausgewählten Knoten eng beieinanderliegen und ein neuer naher Knoten keinen weiteren positiven Effekt bringt.

3.Allgemeinere Modelle

Im folgenden seien zwei allgemeine Modell beschrieben.

Allgemeines Treshold-Modell

f_v bildet Teilmenge der Nachbarn eines Knotens v auf [0,1] ab
$f_v(\emptyset)=0$
Jeder Knoten wählt am Anfang des Prozesses zufällig Θ_v aus [0,1]
v ist in Schritt t aktiv, wenn $f_v(S)\geq\Theta_v$,
S ist dabei die Menge der in t-1 aktive Nachbarn.

Allgemeines Cascade-Modell:

Beim allgemeinen Cascade-Modell ist Wahrscheinlichkeit der Aktivierung nicht nur von dem Knoten, der eine Aktivierung versucht abhängig, sondern auch von allen Knoten, die beim Aktivierungsversuch gescheitert sind.
Beide Modelle sind äquivalent, aber ihre Approximation ist NP-schwer. Es gibt allerdings auch allgemeinere Modelle, die noch Approximierbar sind. Beim *Triggering Set*-Modell zum Beispiel wählt jeder Knoten zufällig und unabhängig einen Knotenmenge den *Triggering Set* aus seinen Nachbarn aus. Ein Knoten wird dann in Schritt t aktiv, wenn ein

Knoten aus seinem *Triggering Set* in t-1 aktiv ist.

4.Nichtprogressive Prozesse:

Um das Modell zu erweitern soll es möglich sein Knoten abzuschalten und in einen laufenden Prozess einzugreifen.
Beim Linear Treshold Modell wählt ein Knoten wählt in jedem Schritt seinen Grenzwert Θ per Zufall neu. Knoten v ist in Schritt t aktiv, wenn $f_v(S) \geq \Theta_v$. Eine **Intervention** ist die Aktivierung eines bestimmten Knotens zu zu einem bestimmten Zeitpunkt t. Das *Influence Maximization Problem* ist dann im nichtprogressiven Fall für eine Zahl k, die Suche nach k-Interventionen, die den Einfluss maximieren. Nichtprogressive Prozesse lassen sich über einen äquivalent über ein progressives Modell mit einem mehrschichtigen Graphen darstellen und sind approximierbar.

Allgemeine Marketingstrategien:

Die bisherige Annahme, dass man Knoten direkt aktivieren kann ist eine sehr vereinfachte Sichtweise. Ein realistischeres Modell ist die Einführung von Marketinghandlungen. Es existieren eine Anzahl m an Marketinghandlungen M_i. Jedes M_i kann die Wahrscheinlichkeit von einem oder mehreren Knoten aktiviert zu werden beeinflussen. Das Investment x_i gibt an wieviel Investitionen in Form von Geld in die Marketing Aktionen M_i geflossen sind. Eine Markting Strategie ist dann ein m-dimensionaler Vektor x von Investionen. Die zu maximierende Funktion ist dann:

$$g(x) = \sum_{A \subseteq V} \sigma(A) \cdot \prod_{u \in A} h_u(x) \cdot \prod_{u \notin A} (1 - h_v(x))$$

Wobei $h_v(x)$ die Wahrscheinlichkeit beschreibt, dass Knoten v aktiviert wird.
g(x) ist dabei mittels Hill-Climbing approximierbar.

Quellen

[1]David Kempe, Jon Kleinberg, Éva Tardos:„Maximizing the Spread of Influence trough a sozial network“