ЭЛЕКТРОННЫЕ БИБЛИОТЕКИ:
ПЕРСПЕКТИВНЫЕ МЕТОДЫ И ТЕХНОЛОГИИ,
ЭЛЕКТРОННЫЕ КОЛЛЕКЦИИ
XI Всероссийская научная конференция

English version

Петрозаводск,
17 - 21 сентября, 2009

Метод выявления неявных связей объектов

Снарский А.А., Ландэ Д.В., Женировский М. И.

НТУУ «Киевский политехнический институт»,
Информационный центр «ЭЛВИСТИ»,
Институт теоретической физики им. Н.Н. Боголюбова НАН Украины

ПРЕДМЕТНАЯ ОБЛАСТЬ

В настоящее время в теории и практике аналитической деятельности получила большое развитие концепция сложных сетей, являющаяся с одной стороны, развитием теории графов, а с другой стороны, областью применения подходов, применяемых в физике, например, в теории электрических цепей или теории перколяции. Переход к физической парадигме объясняется, по-видимому, именно сложностью сетей, которые, на самом деле окружают нас повсюду. В частности, сети, образуемые персонами, совместно упоминаемыми в одних и тех же публикациях, позволяют аналитикам делать выводы об общих интересах отдельных групп персон, выявлять неявные связи, пренебрегать несущественными и т.п.

Описывается метод, позволяющий выявлять неявные связи в сложных сетях, представленных матрицами инцидентности.

Описывается применение данного метода, базирующегося на теории электрических сетей, для выявления силы взаимосвязей понятий, извлекаемых из неструктурированных текстов, в частности, персон.

Этот же метод может применяться, например, для выявления неявных связей терминов в текстах сообщений электронных СМИ.

ТРАДИЦИОННЫЕ ПОДХОДЫ

Известно, что матрицы взаимосвязей понятий (МВП) являются одной из форм представления сетевых структур, аналогичной по функциональности их графовому представлению. На практике эти матрицы чаще всего отражают близость отдельных понятий (совместную встречаемость в документах или близость по сопутствующему контексту в разных документах). При самых различных подходах к их построению - это, как правило, симметричные матрицы, элементы которых – коэффициенты взаимосвязей. Если отношения между понятиями не носят направленного характера, то их также можно рассматривать как неориентированные графы и применять к ним соответствующие методы.

Чаще всего ребрам этих графов приписываются весовые коэффициенты, которые пропорциональны количеству документов из некоторого массива, одновременно соответствующие обоим узлам (понятиям), соединяемым этими ребрами. Существуют и другие многочисленные подходы к определению близости понятий в массивах неструктурированных текстов, среди таких можно назвать контекстные, вероятностные и энтропийные (Mutual Information), но все они являются лишь предпосылками для построения матриц взаимосвязей, их перегруппировки и визуализации.

Таблица взаимосвязи понятий

Обозначим $p_i \ (i = 1,\ldots,K)$ - понятие, $d^j \ (j = 1,\ldots,N)$ - документ, $d^j \in D$ - массив документов, e_i^j - признак соответствия понятия p_i документу d^j:

$$e_i^j = \begin{cases} 1, & p_i \in d^j \\ 0, & p_i \notin d^j \end{cases}$$

Можно определить уровень связи понятий p_i и p_k:

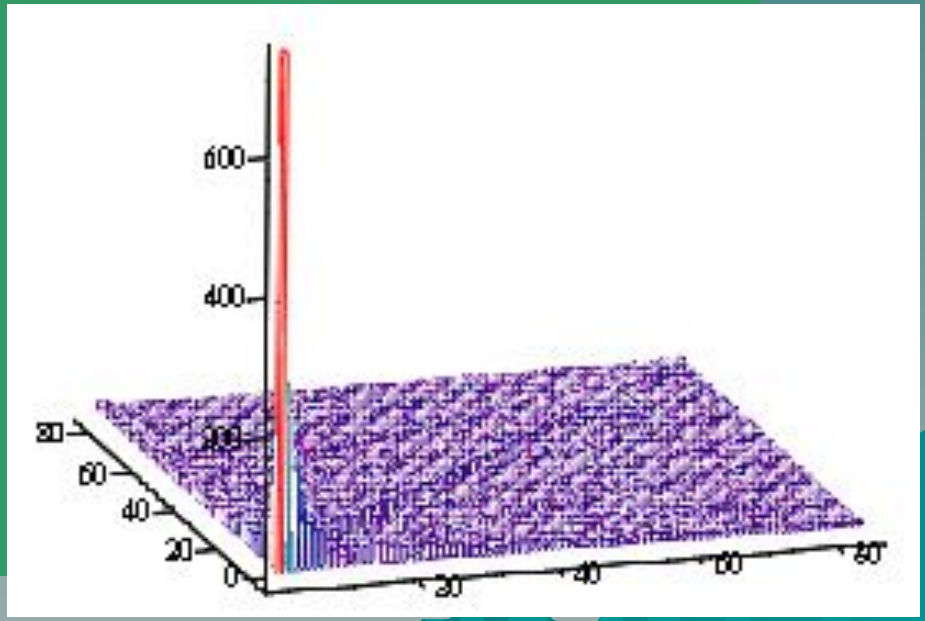
$$M_{ik} = \sum_{j=1}^{N} e_i^j e_k^j .$$

Введя обозначение: $E = \left\| e_i^j \right\|_{i=1,..,K}^{j=1,..,N}$,

получаем:

$$M = EE^T = \left\| M_{ik} \right\|_{i,k=1,..,K}.$$

Недиагональный элемент $M_{ik} \ (i \neq k)$ равен количеству одновременных упоминаний узлов (персон) i и k во всех статьях из базы данных.



Коэффициент сцепления

Будем трактовать значение элемента матрицы M - M_{ik}, как числа, которое приписывается весу связи (ребра) между i и k, в качестве проводимости этой связи, по аналогии с теорией электрических цепей. Тогда по аналогии с этой теорией можно ввести вспомогательную матрицу A:

$$A_{ik} = -M_{ik},$$

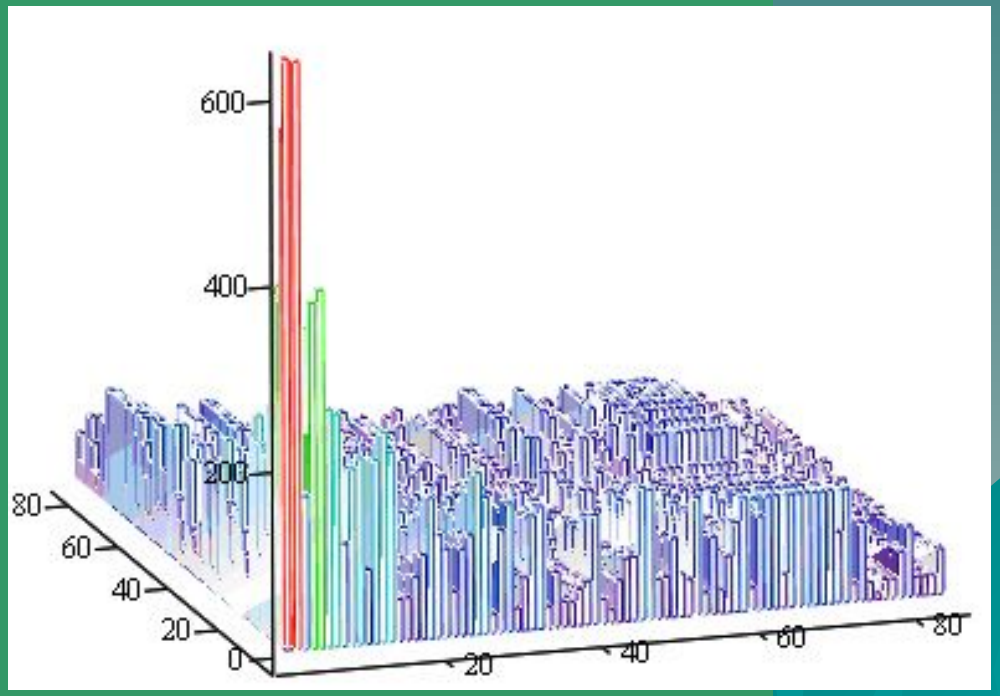
$$A_{ii} = \sum_{j \neq i} \left| A_{ij} \right| .$$

С помощью A можно найти матрицу кондактанса (полной проводимости) G:

$$G_{ik} = \frac{\det(A)}{\det\left(A_{(i+k)(i+k)}\right)} .$$

Здесь $A_{(i+k)(i+k)}$ - это минор матрицы A, который вычисляется следующим образом: строка i прибавляется к строке k и затем вычеркивается, столбец i прибавляется к столбцу k и затем также вычеркивается. Характеристикой всей системы является средний коэффициент сцеплений (когезии) G_{av}, равный

$$G_{av} = \frac{1}{N(N-1)} \sum_{\substack{i,k \\ i \neq k}}^{N} G_{ik} .$$



Неявные связи (матрица скрытности)

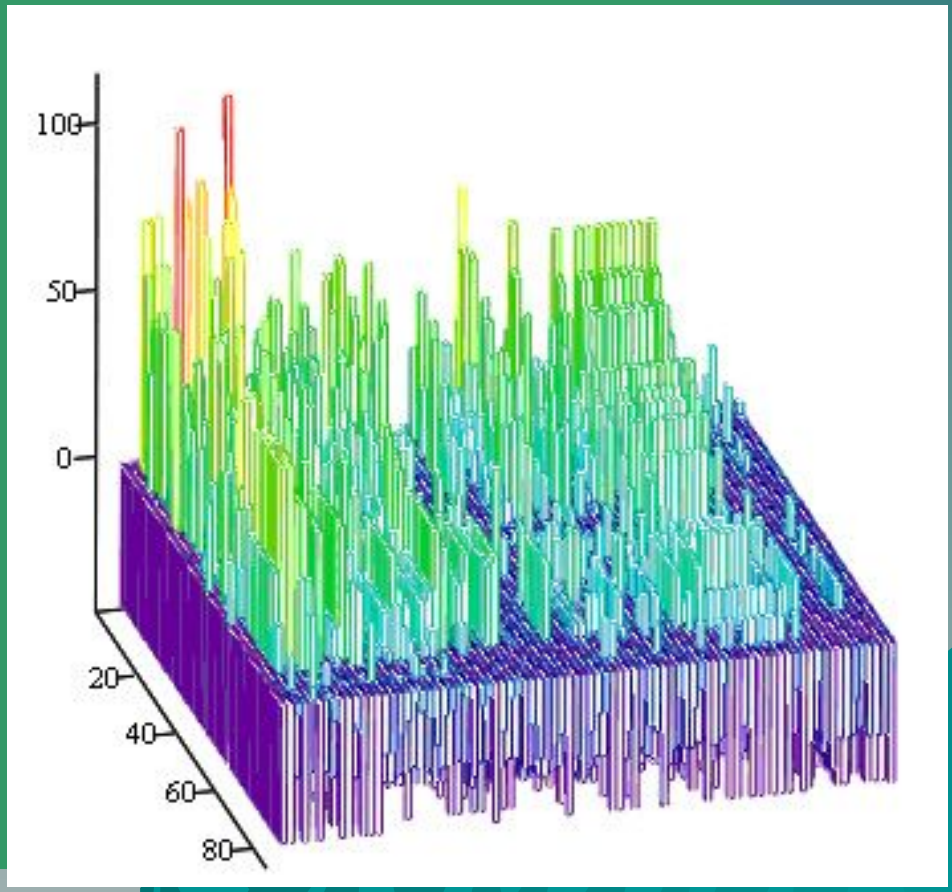
Будем теперь исследовать только непрямые связи между узлами (персонами), условно назовем их скрытыми или неявными связями. Для этого обнулим все значения G_{ik} для тех пар i и k, которые связаны непосредственно (полученную матрицу обозначим как K).

Нас будут интересовать те пары узлов (персон) между которыми нет прямых связей, а коэффициент когезии скрытых связей больше среднего коэффициента когезии всей базы. Для удобного представления последних введем матрицу скрытости F (скрытость - furtive) :

$$F = K - G_{av},$$

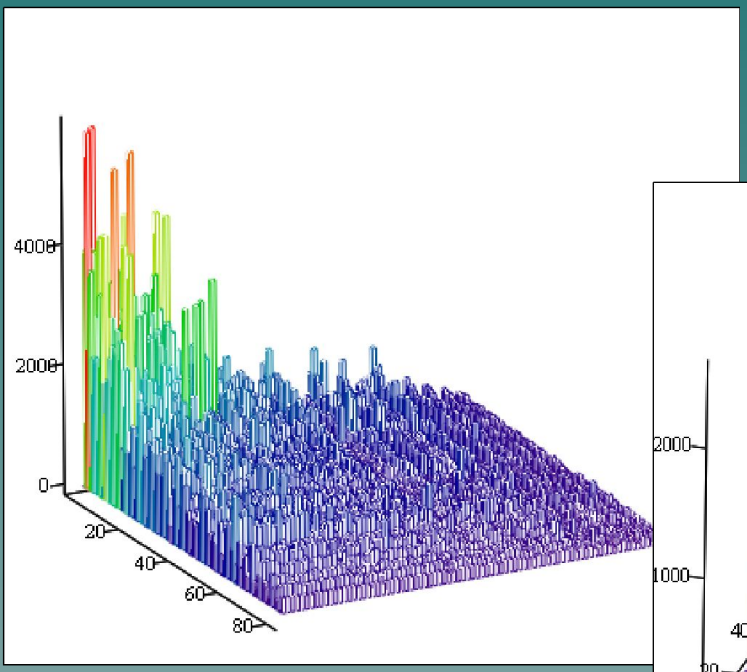
где нас будут интересовать только положительные элементы F.

На рис. изображена матрица скрытости для реальной базы данных персон, для которой была построена матрица взаимосвязей.

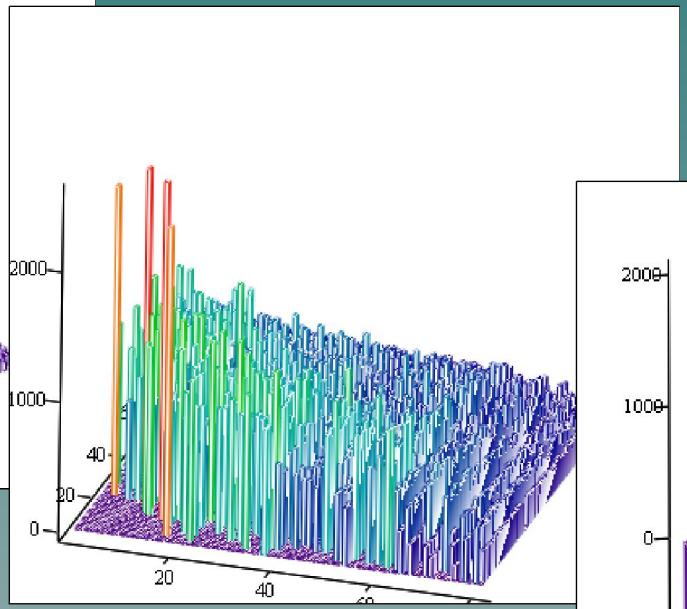


Скрытые связи слов

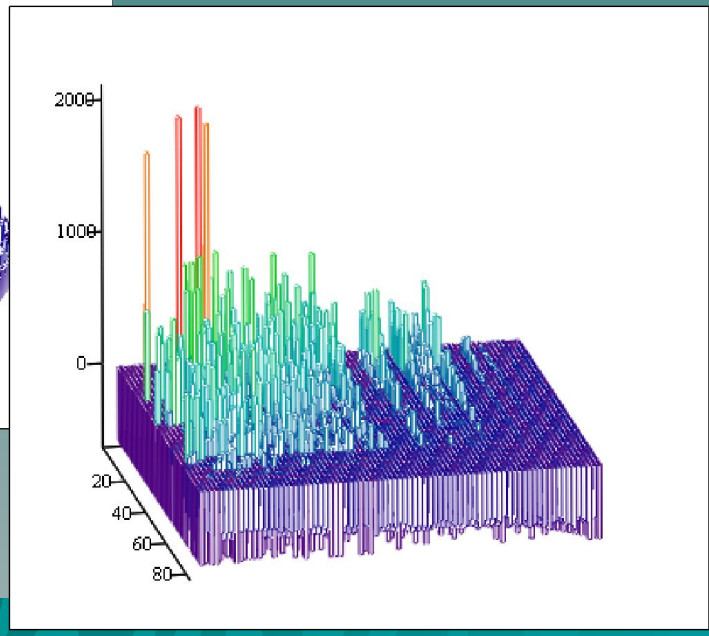
1. Слова считаются связанными, если они стоят рядом с текстом.
2. Известно, что матрица инцидентности слов сильно разряжена.
3. Придуман алгоритм отбора «опорных слов». Выбираются слова, которые участвуют в наиболее часто встречаемых «триадах».



G



G1



G1 – 561.9

Некоторые выводы

Приведенный метод во многом напоминает подходы, базирующиеся на комбинаторном кластерном анализе, однако его принципиальное отличие в том, что он основывается на правилах Кирхгофа о протекании электрического тока в разветвленных цепях. При этом целью было использование методов, уже наработанных в теории электрических сетей.

В отличие от существующих в настоящее время подходов к выявлению взаимосвязей понятий, предложенный метод позволяет выявлять, определять относительный вес и визуализировать неявные связи любых уровней.

Вместе с тем рассмотренное направление анализа сложных сетей сегодня актуально в маркетинговых и социальных исследованиях, в конкурентной разведке, в задачах выявления и визуализации различных сообществ.

Спасибо за внимание!

Д.В. Ландэ,
dwl@visti.net

Информационный Центр «ЭЛВИСТИ»,
Киев, Украина