Метод выявления неявных связей объектов

© Снарский А.А.[1], Ландэ Д.В.[1,2], Женировский М. И.[3]

[1]НТУУ «Киевский политехнический институт»,
[2]Информационный центр «ЭЛВИСТИ»,
[3]Институт теоретической физики им. Н.Н. Боголюбова НАН Украины
asnarskii@gmail.com, dwl@visti.net

Аннотация

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.

1 Матрицы инцидентности

В настоящее время в теории и практике аналитической деятельности получила большое развитие концепция сложных сетей (complex networks) [16], являющаяся с одной стороны, развитием теории графов, а с другой стороны, областью применения подходов, применяемых в физической науке, например, в теории электрических цепей или теории перколяции. Переход к физической парадигме объясняется, по-видимому, именно сложностью этих сетей, которые, на самом деле окружают нас повсюду – это и транспортные сети, и сети цитирования, и, безусловно, Интернет [8]. В частности, сети, образуемые персонами, совместно упоминаемыми в одних и тех же публикациях, позволяют аналитикам делать выводы об общих интересах отдельных групп персон во времени [15], выявлять неявные связи [10], пренебрегать несущественными и т.п.

Технологиям выявления понятий из неструктурированных текстов посвящено достаточно много публикаций [1,2,12], эта проблематика выходит за рамки нашего исследования. В данной работе предлагается метод исследования сети понятий, характеризующейся большим количеством узлов, ребер (связей) с различными весовыми значениями, высокой динамикой появления новых узлов и связей. Известно, что матрицы взаимосвязей понятий (МВП) [5,6] 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) [5,9,13], но все они являются лишь предпосылками для построения матриц взаимосвязей, их перегруппировки и визуализации [11,14].

Рассмотрим одно из формальных определений матрицы взаимосвязей понятий M, соответствующее приведенным в работах [5] и [6]. Обозначим p_i $(i = 1,...,K)$ – понятие, d^j $(j = 1,...,N)$ – 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:

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

Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.

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

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

Будем называть данную матрицу инцидентности M матрицей взаимосвязей понятий. Недиагональный элемент M_{ik} $(i \neq k)$ этой матрицы равен количеству одновременных упоминаний узлов (персон) i и k во всех статьях из базы данных. Диагональный элемент матрицы M_{ii} - это количество упоминаний i - того узла (персоны) во всех документах. На рис. 1 приведен пример трехмерного изображения матрицы взаимосвязи понятий, состоящей из 84 узлов (персон). Данная матрица была получена на основании анализа массива веб-публикаций, сосканированных системой контент-мониторинга InfoStream [3] в течение первого квартала 2009 года по тематике деятельности Киевской городской государственной администрации. Объем исходных данных составил свыше 10 тыс. документов.

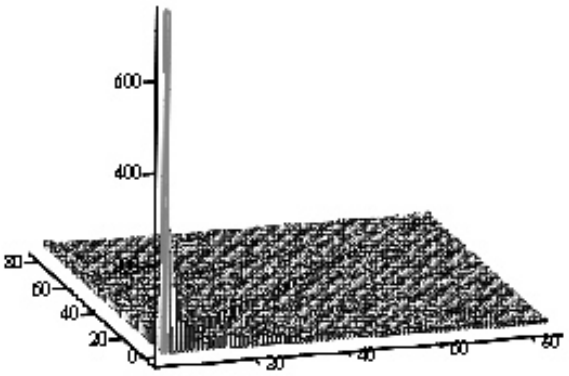

Рис.1.Изображение модифицированной матрицей связей M. По горизонтальным осям отложены номера узлов (персон), по вертикальной – весовые значения связей

2 Коэффициент сцепления

В рамках рассматриваемого метода предлагается новая характеристика сложных сетей - коэффициент сцепления (cohesion). Пусть мы имеем некоторую сеть, узлами которой являются люди (персоны), а ребрами - некоторые отношения между ними (как такие отношения, можно рассматривать общие интересы, упоминаемость в одних и тех же документах, и т.п.). При этом каждый из узлов связан с некоторым количеством других узлов этой сети. Актуальной является задача исследования такой сети - выяснение, какие узлы в ней играет ведущую роль и, главное, насколько эти главные узлы хорошо связаны (сцеплены), между собой. То есть на входе имеется стандартная матрица инцидентности M, соответствующая исходному графу связей, а на выходе хотим получить номера так называемых главных узлов и узнать, насколько хорошо эти узлы сцеплены между собой.

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 (см., например, [4, 7]). Тогда по аналогии с этой теорией можно ввести так называемую матрицу инцидентности проводимости A для матрицы M:

$$A_{ik} = -M_{ik},$$
$$A_{ii} = \sum_{j \neq i} \left| A_{ij} \right|$$

Здесь A_{ii} - сумма проводимостей ребер, инцидентных данному узлу, а A_{ik} - проводимость прямой связи между узлами (персонами) i и k взятая со знаком минус.

Зная матрицу A можно найти матрицу кондактанса (полной проводимости) G, каждый элемент которой G_{ik} соответствует полной проводимости с учетом всех прямых и непрямых связей между двумя узлами i и k ($i \neq k$). Будем называть величину G_{ik} коэффициентом когезии (сцепления):

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

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. Если один из индексов равен нулю, то просто вычеркивается столбец и строка, соответствующие ненулевому индексу.

Для реальной базы данных персон, для которой построена матрица взаимосвязей, полученная матрица G графически представлена на рис. 2

Характеристикой всей системы является средний коэффициент сцеплений (когезии) G_{av}, равный

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

Для пояснения смысла вводимых параметров рассмотрим также «игрушечную» небольшую базу данных связей (сеть) из четырех узлов (персон), для которой матрица M имеет вид:

$$M^1 = \begin{pmatrix} 5 & 2 & 3 & 0 \\ 2 & 3 & 0 & 1 \\ 3 & 0 & 6 & 3 \\ 0 & 1 & 3 & 4 \end{pmatrix}$$

На рис. 3 эта же база данных изображена в виде графа, около каждой связи, которой проставлен значение проводимости (совместных упоминаний).

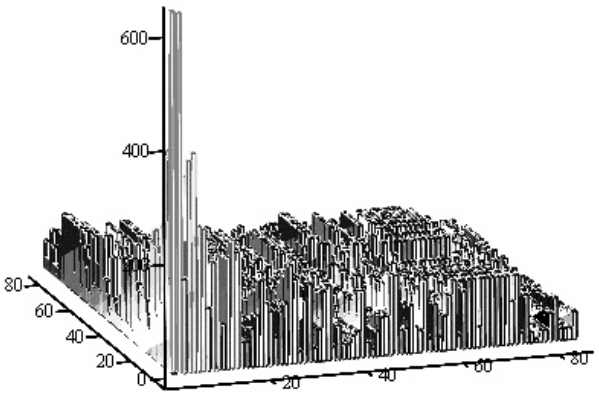

Рис. 2. Матрица когезии. По горизонтальным осям отложены номера узлов (персон), по вертикальной – значения матрицы

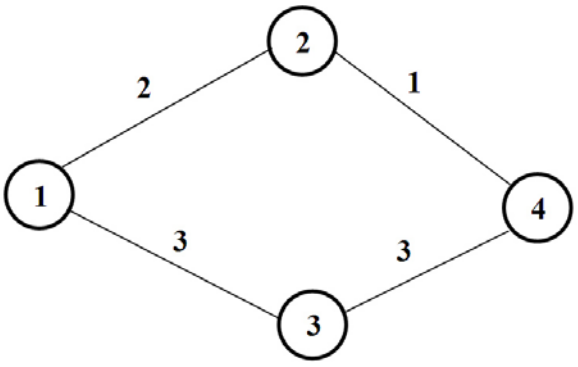

Рис. 3. Граф модифицированной матрицы связей из четырех узлов (персон) M^1

Матрица сцепления (когезии) G^1, соответствующая матрице M^1, равна:

$$G^1 = \begin{pmatrix} - & 2.6 & 3.6 & 2.2 \\ 2.6 & - & 2 & 1.9 \\ 3.6 & 2 & - & 3.5 \\ 2.2 & 1.9 & 3.5 & - \end{pmatrix},$$

а среднее значение $G^1_{av} = 2.2$.

Как видно из этого примера матрица когезии, в отличие от матриц A и M учитывает (с соответствующим весом) все, а не только прямые связи. В самом деле, элемент $M^1_{1,4}$ равен нулю, между узлами 1 и 4 нет прямой связи. В тоже время, между этими узлами есть опосредованные связи через узлы 2 и 3. Заметим, что коэффициент когезии $G^1_{1,4}$ узлов 1 и 4, непосредственно не связанных между собой больше чем этот же коэффициент $G^1_{2,4}$ для связанных между собой узлов 2 и 4.

3. Применение

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) :

$F = K - G_v$,

где нас будут интересовать только положительные элементы F.

На рис. 4 изображена матрица скрытости для реальной базы данных персон, для которой была построена матрица взаимосвязей.

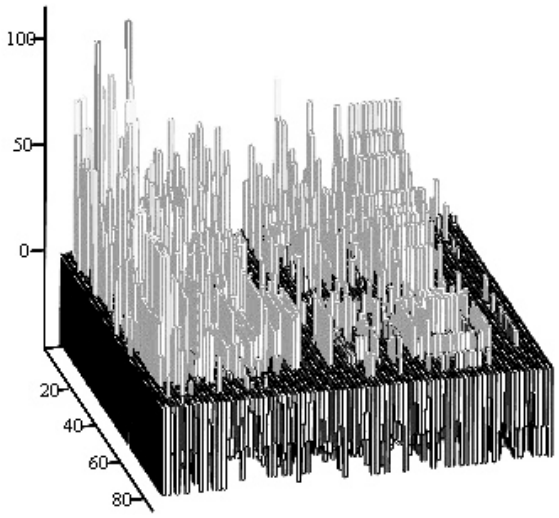

Рис. 4. Матрица скрытости F. По вертикальной оси показан коэффициент когезии, из которого вычтено среднее значение когезии по всей базе, тех пар узлов, между которыми нет непосредственной связи

Рассмотрим несколько конкретных узлов (персон) из МВП (см. Табл.). Для выбранных персон значения матрицы A равны: $A_{3,3} = 170$, $A_{4,4} = 526$, $A_{9,9} = 234$, $A_{12,12} = 242$, $A_{27,27} = 20$, откуда в частности следует, что максимальное число связей у Ю. Тимошенко. А значения матрицы M равны: $M_{3,4} = 0$, $M_{3,9} = 0$, $M_{3,12} = 0$, $M_{3,27} = 0$, $M_{4,9} = 18$, $M_{4,12} = 26$, $M_{4,27} = 0$, $M_{9,12} = 18$, $M_{9,27} = 0$, $M_{12,27} = 0$. Приведем также коэффициент сцепления для узлов 3 и 9 ($G_{3,9} = 89.4$). Анализируя полученные значения,

можно, в частности, заметить, что между узлами 3 и 9 нет прямой связи. В тоже время коэффициент сцепления равен 89.4, что более чем в два раза выше среднего коэффициента сцепления по МВП.

Табл. Несколько узлов сети персон

Номер узла	Фамилия человека, соответствующая данному номеру
3	Басс Д.
4	Тимошенко Ю.
9	Яценюк А.
12	Янукович В.
27	Кильчицкая И.

Приведенный метод во многом напоминает подходы, базирующиеся на комбинаторном кластерном анализе, однако его принципиальное отличие в том, что он основывается на законах Максвелла, из которого, в частности, следуют известные закономерности Кирхгофа о протекании электрического тока в разветвленных цепях. При этом не ставилась задача поиска прямых аналогий, а скорее 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. Следует отметить, что аналоги подобных методов из теории электрических цепей, до сих пор не находили широкого применения в практике аналитической обработки информации.

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.

Литература

[1] Гаврилова Т.А., Червинская К.Р. Извлечение и структурирование знаний для экспертных систем. - М.: Радио и связь, 1992.

[2] Гершензон Л. М., Ножов И.М. , Панкратов Д. В. Система извлечения и поиска структурированной информации из больших текстовых массивов СМИ. Архитектурные и лингвистические особенности // Компьютерная лингвистика и интеллектуальные технологии: труды Международного семинара Диалог'2005. – М.: Наука, 2005.

[3] Григорьев А.Н., Ландэ Д.В. и др. InfoStream. Мониторинг новостей из Интернет: технология, система, сервис: научно-методическое пособие. – Киев: ООО «Старт-98», 2007. – 40 с.

[4] Джексон Дж. Классическая электродинамика - М., Мир, 1965. - 694 с.

[5] Додонов А.Г., Ландэ Д.В. Выявление понятий и их взаимосвязей в рамках технологии контент-мониторинга // Регистрация, хранение и обработка данных, 2006, Т. 8, № 4.– С. 45 - 52.

[6] Калиткин Н.Н., Карпенко Н.В., Михайлов А.П. и др. Математические модели природы и общества –М.: Физматлит, 2005. -360 с.

[7] Попов В.П. Основы теории цепей – М.: Высшая школа, 1985. - 496 с.

[8] Albert R., Jeong H., Barabasi A. Attack and error tolerance of complex networks // Nature. - 2000. - Vol. 406. - pp. 378-382.

[9] Church K.W. , Hanks P. Word association norms, mutual information, and lexicography, Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics, 1989.

[10] Clauset A., Moore C., Newman M. E. G. Hierarchical structure and the prediction of missing links in networks // Nature. - 2000. - Vol 453. - pp. 98–101.

[11] Danon L. , Dнaz-Guilera A. , Duch J., Arenas A.. Comparing community structure identification // J. Stat. Mech. (2005) P09008. doi: 10.1088/1742-5468/2005/09/P09008 PII: S1742-5468 (05) 07477-7.

[12] Grishman R. Information extraction: Techniques and challenges. In Information Extraction (International Summer School SCIE-97). Springer-Verlag, 1997.

[13] Guiasu, S. Information Theory with Applications, McGraw-Hill, New York, 1977.

[14] Knepper M.M. , Killam R., Fox K.L., Frieder O. Information Retrieval and Visualization using SENTINEL / TREC 1998: 336-340.

[15] Lande D.V., Snarskii A.A. Dynamic network of concepts from web publications // ePrint Arxiv (0806.1439).

[16] Newman M.E.J. The structure and function of complex networks // SIAM Review. - 2003. - Vol. 45. - pp. 167-256.

Discovering implicit relations of concepts

A.A. Snarskii, D.V. Lande, M.I. Zhenirovsky, NTUU "KPI", ElVisti IC, Bogolyubov Institute for Theoretical Physics, Kyiv, Ukraine

The method of discovering implicit relations in complex networks presented by incidence matrixes is described, along with implementation of this method based on electric circuit theory, for relevation of concepts correlations, e.g. persons, derived from unstructured texts.