ВЗАИМОСВЯЗЬ ПОНЯТИЙ В ДОКУМЕНТАХ – СОВМЕСТНОЕ ПОЯВЛЕНИЕ ИЛИ КОНТЕКСТНАЯ БЛИЗОСТЬ?

INTERRELATION OF DOCUMENTS CONCEPTS - JOINT OCCURRENCE OR CONTEXTUAL AFFINITY?

Ландэ Д.В. , dwl@visti.net, Григорьев А.Н. , gri@visti.net, Дармохвал А.Т., hval@visti.net,
Информационный центр «ЭЛВИСТИ», Киев

Приведены решения, позволяющие выявлять силу взаимосвязей понятий, извлекаемых из неструктурированных текстов, на основе применения двух алгоритмов, первый из которых основывается на учете совместного вхождения этих понятий в одни и те же документы, а второй на учете общего контекстного окружения. Рассматриваются два вида таблиц взаимосвязей понятий. Таблицы первого вида всегда отражают взаимосвязи понятий точнее, а второго - более полно.

Перспективным направлением развития технологии интеграции информационных ресурсов [1] является автоматическое извлечение понятий из неструктурированных текстов, а также выявление их взаимосвязей.

Технологиям выявления фактографии из неструктурированных текстов посвящено достаточно много публикаций [2-5], подход авторов близок к описанному в [3]. Однако, предметом данного доклада является не выявление понятий, а сравнение двух из множества существующих подходов к построению таблиц взаимосвязей понятий. Известно, что таблицы взаимосвязей понятий [6, 7] строятся как статистические отчеты, отражающие близость отдельных понятий (совместную встречаемость в документах или близость по сопутствующему контексту в разных документах). Это, как правило, симметричные матрицы, элементы которых – коэффициенты взаимосвязей, соответствующие ее строкам и столбцам. Эти матрицы можно также рассматривать как неориентированные графы и применять к ним соответствующие методы. Как правило, узлы этих графов – коэффициенты, которые пропорциональны количеству документов из некоторого массива, одновременно соответствующие обоим понятиям, или количеству других понятий, употребляемых совместно с данными понятиями. Таким образом взаимосвязь понятий может быть оценена с помощью двух алгоритмов:

- совместного вхождения – путем расчета совместного вхождения этих понятий в одни и те же документы;
- контекстной близости - путем расчета корреляций наборов смежных понятий, которые входят в документы, в которых упоминались данные понятия.

Существуют и некоторые другие подходы к определению близости терминов в массивах неструктурированных текстов, в частности, вероятностные или энтропийные (Mutual Information) [8, 9], но все они являются лишь предпосылками для построения таблиц взаимосвязей, их перегруппировки и визуализации [10-13].

Рассмотрим формальное определение таблицы взаимосвязей понятий *TVP'*, построенной с помощью первого из приведенных выше алгоритмов. Обозначим p_j – понятие ($j=1,..,M$), D_i – документ ($i=1,\ldots,N$), $D_i \in D$ – массив документов, e_{ij} – признак соответствия понятия документу:

$$p_j \in D_i \Rightarrow e_{ij} = 1, \text{ иначе } e_{ij} = 0.$$

Можно определить уровень связи понятий p_j и p_k:

$$v'_{jk} = \sum_{i=1}^{N} e_{ji} e_{ki}$$

Введя обозначение: $E = \| e_{ij} \|_{j=1,\ldots,M;\ i=1,\ldots,N}$, получаем:

$$TVP' = E^T E = \| v'_{jk} \|_{j,k=1,\ldots,M} .$$

Для случая второго алгоритма, учитывающего контекстную близость (множество понятий, входящих в документы одновременно с заданными), определим таблицу взаимосвязей понятий *TVP''* следующим образом. Обозначим $W_i = \{p_1,\ldots,p_L\}$ – множество понятий из D_i.

Рассмотрим множество понятий, содержащихся в тех же документах из массива *D*, что и понятие p_j :

$$IP(p_j) = \bigcup_{\{D_i \in D\}} W_i$$

Рассмотрим также матрицу $T(p_j)$ с элементами t_{ij}, соответствующие $IP(p_j)$:

$$p_i \in IP(p_j) \Rightarrow t_{ji} = 1, \text{ иначе } t_{ij} = 0;$$

$$T(p_j) = \| t_{ij} \|_{i=1,\ldots,M} .$$

В этом случае уровень связи понятий p_j и p_k можно определить следующим образом:

$$v''_{jk} = (T(p_j), T(p_k)) = \sum_{i=1}^{M} t_{ij} t_{ik}$$

Таким образом, таблица взаимосвязей понятий будет иметь вид:

$$TVP'' = \| v''_{jk} \|_{j,k=1,\ldots,M} .$$

Основное отличие двух таблиц взаимосвязей (рис. 1) заключается в том, что таблица взаимосвязей первого вида всегда отражает взаимосвязи понятий точнее, чем таблица взаимосвязей второго типа, однако, таблица второго типа учитывает взаимосвязи более полно ($v'_{jk} > 0 \Rightarrow v''_{jk} > 0$, действительно, $v'_{jk} > 0 \Rightarrow \exists i : p_j \in D_i, p_k \in D_i \Rightarrow (T(p_j), T(p_k)) = v''_{jk} > 0$).

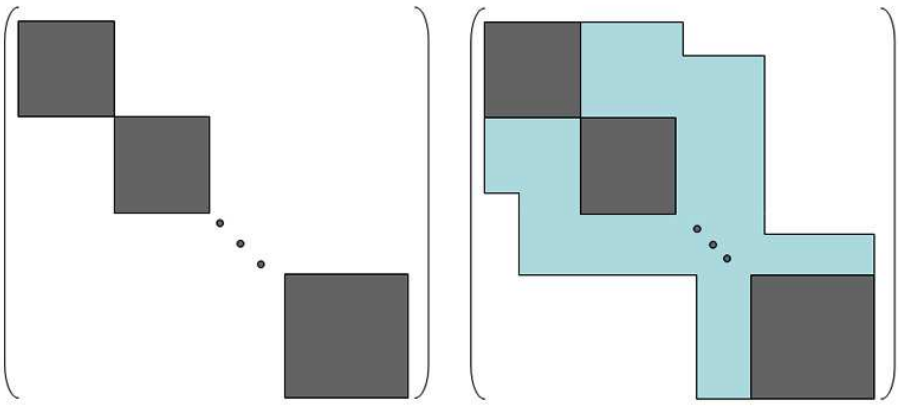

Рис. 1. Два варианта таблицы взаимосвязей понятий

Обратное утверждение в общем случае неверно. Проведем мысленный эксперимент, подтверждающий это замечание. Рассмотрим два понятия «пингвин» и «белый медведь». Эти понятия могут иметь ненулевое контекстное пересечение за счет таких ключевых слов, как «лед», «мороз», «рыба», однако понятие «пингвин» входит в документы, описывающие фауну Антарктики, а «белый медведь» - фауну Арктики.

Для переупорядочения понятий из таблицы взаимосвязей с целью выявления блоков – множеств наиболее взаимозависимых понятий (рис. 2) в рамках системы контент-мониторинга InfoStream [14]

авторами применялись алгоритмы кластерного анализа, в частности, k-means, который является одним из самых эффективных для группировки динамических данных [15].

Однако задача оптимальной группировки векторов в данном случае усложняется необходимостью при перестановке номеров векторов-строк одновременно переставлять соответствующие их компоненты.

На рис. 3 представлена трехмерная визуализация первого и второго алгоритма построения таблиц взаимосвязей понятий (график, соответствующий первому алгоритму, для наглядности приподнят на 200 пунктов).

Следует отметить, что в качестве понятий в контексте данного исследования рассматривались наименования компаний, географические названия, персоны, ключевые слова.

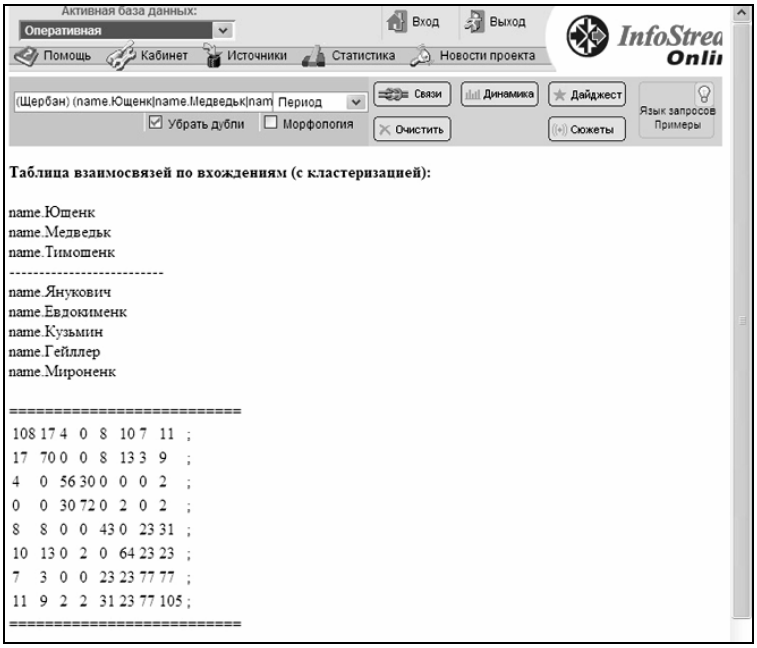

Рис. 2. Кластеризация таблицы взаимосвязей понятий в системе InfoStream [14]

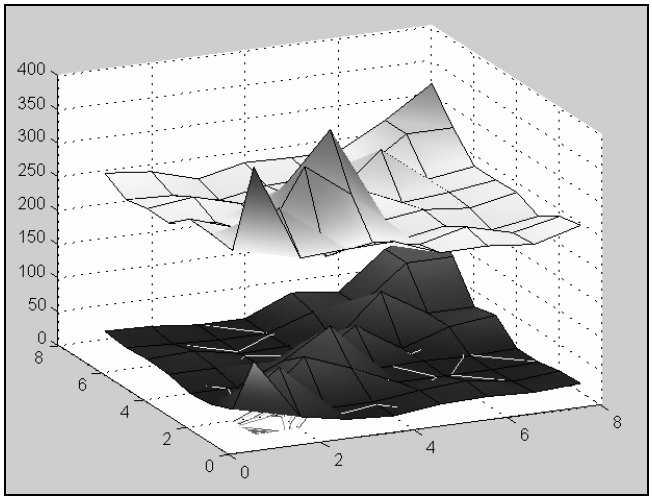

Рис. 3. Трехмерное представление взаимосвязей понятий

Авторам известно несколько разработок в направлении выявления взаимосвязей понятий, извлекаемых из неструктурированных текстов. Сегодня это направление особо актуально в маркетинговых и социальных исследованиях, в задачах выявления и визуализации различных сообществ, которые широко применяются в информационно-аналитических системах поддержки принятия решений (ППР) самых разных уровней. Описанные подходы к построению таблиц взаимосвязей как первого, так и второго видов были реализованы

авторами при проектировании систем ППР на основе технологии InfoStream, которые доступны аналитикам. Предпочтения при использовании определяются ситуативно, в зависимости от того, что более востребовано в текущей задаче, полнота или точность.

Вместе с тем, развитие направления несколько сдерживается недостаточными теоретическими результатами. В частности, своего решения ждут проблемы выявления взаимосвязей с учетом некоторых дополнительных семантических характеристик, в простейшем случае - определение принадлежностей взаимосвязей к положительным (группирующим) или отрицательным (антагонистическим).

Список литературы

1. Ландэ Д.В. Основы интеграции информационных потоков - Киев: Инжиниринг, 2006. — 240 с.
2. R. Grishman. Information extraction: Techniques and challenges. In Information Extraction (International Summer School SCIE-97). Springer-Verlag, 1997.
3. Л. М. Гершензон, И.М. Ножов, Д. В. Панкратов. 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.
4. Протасов С. Обучение с нуля грамматики связей для русского языка // Десятая национальная конференция по искусственному интеллекту с международным участием КИИ-2006. –М.: Наука, 2006.
5. Гаврилова Т.А., Червинская К.Р. Извлечение и структурирование знаний для экспертных систем. - М.: Радио и связь, 1992.
6. Калиткин Н.Н., Карпенко Н.В., Михайлов А.П. и др. Математические модели природы и общества – М.: ФИЗМАТЛИТ, 2005. -360 с.
7. 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.
8. K.W. Church, P. Hanks. Word association norms, mutual information, and lexicography, Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics, 1989.
9. Guiasu, S. Information Theory with Applications, McGraw-Hill, New York, 1977.
10. J.P. Bagrow, E.M. Bollt. Local method for detecting communities // Physical Review E, 2005.
11. L. Danon, A. Díaz-Guilera, J. Duch, A.Arenas. Comparing community structure identification // J. Stat. Mech. (2005) P09008. doi:10.1088/1742-5468/2005/09/P09008 PII: S1742-5468(05)07477-7.
12. M.M. Knepper, R. Killam, K.L. Fox O. Frieder. Information Retrieval and Visualization using SENTINEL / TREC 1998: 336-340.
13. Григорьев А.Н., Ландэ Д.В. Многоуровневый классификатор-навигатор по откликам информационно-поисковой системы // Компьютерная лингвистика и интеллектуальные технологии: труды международной конференции Диалог'2006 – М.: Наука, 2006. - С. 329-331.
14. Григорьев А.Н., Ландэ Д.В., Бороденков С.А., Мазуркевич Р.В., Пацьора В.Н. InfoStream. Мониторинг новостей из Интернет: технология, система, сервис: научно-методическое пособие. – Киев: ООО «Старт-98», 2007. – 40 с.
15. J. B. MacQueen. Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1:281-297, 1967.