TopSOM: визуализация информационных массивов с применением самоорганизующихся тематических карт

В.В. Плешко, вед. разработчик ООО "Гарант-Парк-Интернет"
А.Е. Ермаков, к.т.н., вед. разработчик ООО "Гарант-Парк-Интернет"
Г.В. Липинский, руководитель проекта ООО "Гарант-Парк-Интернет"

Аннотация

Работа посвящена развитию известного метода визуализации массивов полнотекстовых документов WebSOM, основанного на отображении многомерного семантического пространства на плоскость с локальным сохранением топологии, что реализуется с применением самоорганизующихся карт Кохонена. Разработана модификация метода, названная TopSOM, которая использует нейросетевую технологию выявления ключевых тем документа и позволяет формировать тематические карты информационных массивов с более высоким качеством, чем исходный вариант метода.

Введение

Стимулированный развитием Internet рост числа полнотекстовых документов, представленных в электронном виде, требует развития соответствующих методов навигации в информационных массивах.

На сегодняшний день основной формой представления текстовой информации следует считать гипертекст, конкретные реализации которого варьируются как по способу построения самого гипертекста (установления связей), так и по форме визуального отображения. С наиболее интересными, на наш взгляд, формами, можно познакомиться на сайтах зарубежных фирм, таких как *http://www.links2go.com, http://www.semio.com, http://www.inxight.com*.

В настоящей статье мы расскажем об опыте развития другого известного способа, основанного на отображении многомерного семантического пространства на плоскость, который в изначальной реализации носит название WebSOM (Web Self-Organaizing Maps) и предназначен для представления массивов полнотекстовых документов в виде двумерной карты, раскраска которой отражает плотность распределения образов документов [1].

Конкретные документы при этом связываются со своими областями карты, причем к каждой области может относиться множество близких по содержанию документов - тематический класс. В свою очередь, близким областям обычно соответствуют близкие классы документов, что является основной особенностью карты. Области карты именуются в зависимости от содержания документов, к ним относящихся.

Пользователь выбирает на карте интересующую область и получает класс соответствующих ей документов близкого содержания. Если же ищутся документы, включающие некоторые слова, то результаты поиска также могут быть отражены на карте, что достигается выделением областей, которым принадлежат найденные документы. В итоге пользователь получает возможность оценить тематическое распределение искомой информации.

1. Визуализация данных при помощи самоорганизующихся карт Кохонена

В методе WebSOM можно выделить два этапа: представление документов в виде векторов из некоторого пространства $\Re^n$, и построение плоской карты, отражающей топологию распределения этих векторов в пространстве $\Re^n$.

Для удобства начнем изложение с построения карты.

Пусть документы представлены векторами $x_1,\ldots,x_n$ из Евклидова пространства $\Re^n$. Построение карты осуществляется при помощи самоорганизующейся карты Кохонена (СКК) [2,3]. СКК можно рассматривать как нелинейное отображение пространства $\Re^n$ входящих векторов на узлы плоской решетки, прямоугольной или гексагональной. Обозначим через L число ее узлов. С узлами решетки связаны вектора $m_1,\ldots,m_L$ из $\Re^n$, называемые кодовыми векторами.

Процесс настройки кодовых векторов - обучение СКК - представляет собой процедуру стохастической аппроксимации точек локального максимума плотности распределения кодовых векторов. При этом алгоритм обучения СКК параллельно решает вторую задачу - локальной аппроксимации топологии пространства $\Re^n$ путем размещения близких кодовых векторов в соседних узлах решетке -упорядочение.

Каждому вектору $x \in \Re^n$ ставится в соответствие номер узла c, выбираемый из условия:

$$||x - m_c|| = \min_{1 \le l \le L} \{||x - m_l||\}. \qquad (1)$$

Таким образом, между документами и узлами устанавливается соответствие “многие к одному”. Если вектора $x_1,\ldots,x_N$ адекватно отражают содержание документов, то, в силу свойств отображения Кохонена, все близкие по содержанию документы будут соответствовать одному и тому же или близким узлам решетки.

Процесс настройки кодовых векторов m_l состоит из инициализации и итераций.

Инициализация:

$$m_l(0) = m_{l0},\ l = 1,\ldots,L. \qquad (2)$$

Итерации:

$$m_l(t+1) = m_l(t) + h_{cl}(t)[x(t) - m_l(t)],\ l = 1,\ldots,L,\ t = 0,1,2,\ldots, \qquad (3)$$

где $h_{ci}(t)$ - так называемая, функция соседства, определенная на узлах решетки, c - определяется равенством (1), $x(t)$ - вектор очередного документа из обучающей выборки, предъявляемый на шаге t обучения.

Чаще всего функция соседства задается в виде

$$h_{cl}(t) = \alpha(t) \cdot \exp\left(-\frac{||r_c - r_l||^2}{2\sigma^2(t)}\right), \qquad (4)$$

где r_c и r_l - вектора из $\Re^2$, указывающие на расположение точек c и l на решетке, $\alpha(t)$ - параметр обучения, $\sigma(t)$ - параметр, управляющий количеством соседей. Величины $\alpha(t)$ и $\sigma(t)$ выбираются так, чтобы $0 < \alpha(t) \le 1$, $\sigma(t) > 0$, $\alpha(t) \to 0$ и $\sigma(t) \to 0$ при $t \to \infty$ и $\sum_0^\infty \alpha(t) = \infty$.

Выбор хорошего начального приближения для кодовых векторов m_{i0} на этапе инициализации может сказать большое влияние на число шагов, необходимое для достижения приемлемого качества обучения. Поэтому допустимы различные методы инициализации, как случайный, так и с использованием статистических оценок распределения векторов.

После построения карты производится ее разметка. Если документы обладают дополнительными атрибутами, типа “название темы”, то каждому узлу можно сопоставить наиболее часто встречающееся в нем значение этого атрибута. В противном случае разметка производится экспертом путем анализа содержимого скоплений документов на карте.

2. Выбор векторного представления документов

Остановимся теперь на выборе векторного представления документов.

В исходном варианте метода WebSOM [1] была использована так называемая семантическая сеть Кохонена [4,5].

С каждым узлом решетки связывается набор слов, встретившихся в коллекции документов в схожем контексте, в предположении, что такие слова отражают сходные понятия. Затем для каждого узла подсчитывается, сколько раз в документе встретились слова, ассоциированные с узлом. Полученная гистограмма берется в качестве образа документа. Размерность пространства документов в этом случае оказывается равна числу узлов решетки.

Примеры карт, полученных таким образом авторами WebSOM, можно найти по адресу *http://websom.hut.fi*. Другой пример, представленный на сайте *http://www.neurok.ru*, демонстрирует применение метода WebSOM к русскоязычных документам.

Ознакомившись с указанными демонстрациями, можно убедиться, что использование WebSOM пока не смогло обеспечить качественного представления, способного заинтересовать реального потребителя информационных систем и вывести технологию за рамки лабораторных исследований.

В связи с этим в своих разработках мы использовали другое представление документов, основанное на нейросетевой технологии тематического анализа текста, позволяющей автоматически выявлять ключевые темы, с ранжированием их по значимости [6,7]. При этом в качестве тем выделяются слова и связные словосочетания, входящие в текст документа и наиболее полно характеризующие его содержание. Сохраняя преемственность, мы назвали метод TopSOM – тематические самоорганизующиеся карты.

В этом случае каждому документу ставиться в соответствие тематический вектор x размерности n:

$$x = (\omega_1, \ldots, \omega_n), \qquad (5)$$

где ω_i представляет вес i-ой темы в документе, а n есть общее количество тем, выделенных во всей коллекции документов.

Использование вектора главных тем в качестве образа документа позволяет решить основную проблему выбора адекватных признаков, используемых для классификации, что уменьшает размерность пространства и повышает качество карты. Кроме этого, возникает возможность автоматически маркировать области карты названиями тем.

Эксперименты были проведены в компании “Гарант-Парк-Интернет” (*http://www.metric.ru*) на базе выборки документов из информационной системы “ИС-Парк” по материалам периодической печати.

Так, для коллекции из 10 тысяч документов общее количество тем составило около тысячи. Расчет карты размером 56x40 занял около 10 часов при использовании ПК Pentium-II-400.

Фрагмент построенной карты приведен на рисунке 1.

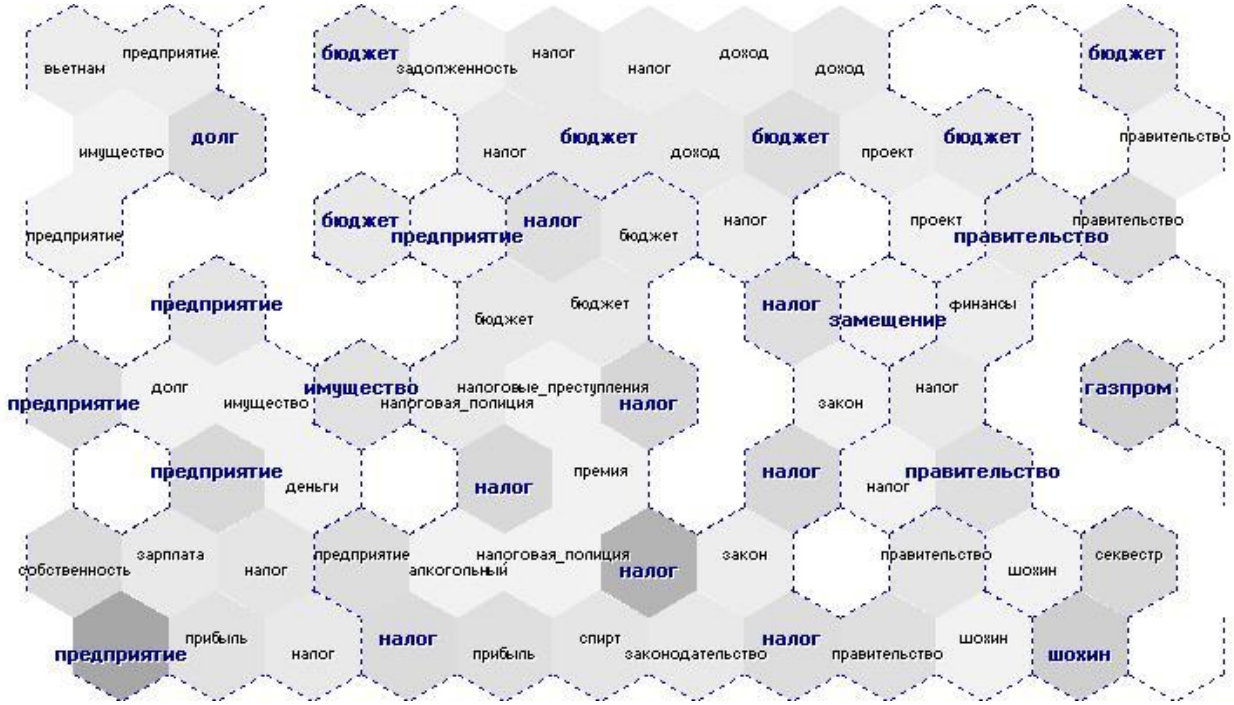

Рис. 1. Фрагмент тематической карты, построенной методом TopSOM

Узлы гексагональной решетки – тематические классы близких документов – здесь представлены шестиугольными областями. Каждой области в общем случае соответствует несколько тем, выделенных в соответствующих ей документах, которые ранжированы по значимости. В силу свойств карты большинство соседних областей образуют группы с одинаковой главной темой, которые обведены пунктиром, а название главной темы написано в некоторых местах выделенным шрифтом. Основные такие группы – это “*долг*”, “*бюджет*”, “*правительство*” “*налог*”,“*предприятие*”. Мелким шрифтом обозначены названия тем, следующих по значимости за главной в группе. Выбранный способ отображения позволяет детализировать информацию. Так, например, документы тематической группы “*предприятие*” в левом нижнем углу дифференцируются на ряд классов, в которых преобладают темы: “*собственность*”, “*зарплата*”, “*прибыль*”, “*налог*”, “*деньги*”, “*имущество*” и “*долг*”.

Ввиду того, что подавляющее большинство документов является политематическим, реальные границы между темами на карте оказываются континуальны. Так, в правом нижнем углу находятся три близких по содержанию класса документов по теме “*Шохин*”, причем документы в одном из них преимущественно относятся к тематической группе “*налог*”, а в другом преобладает тема “*правительство*”.

Яркость окраски пропорциональная количеству документов, относящихся к области.

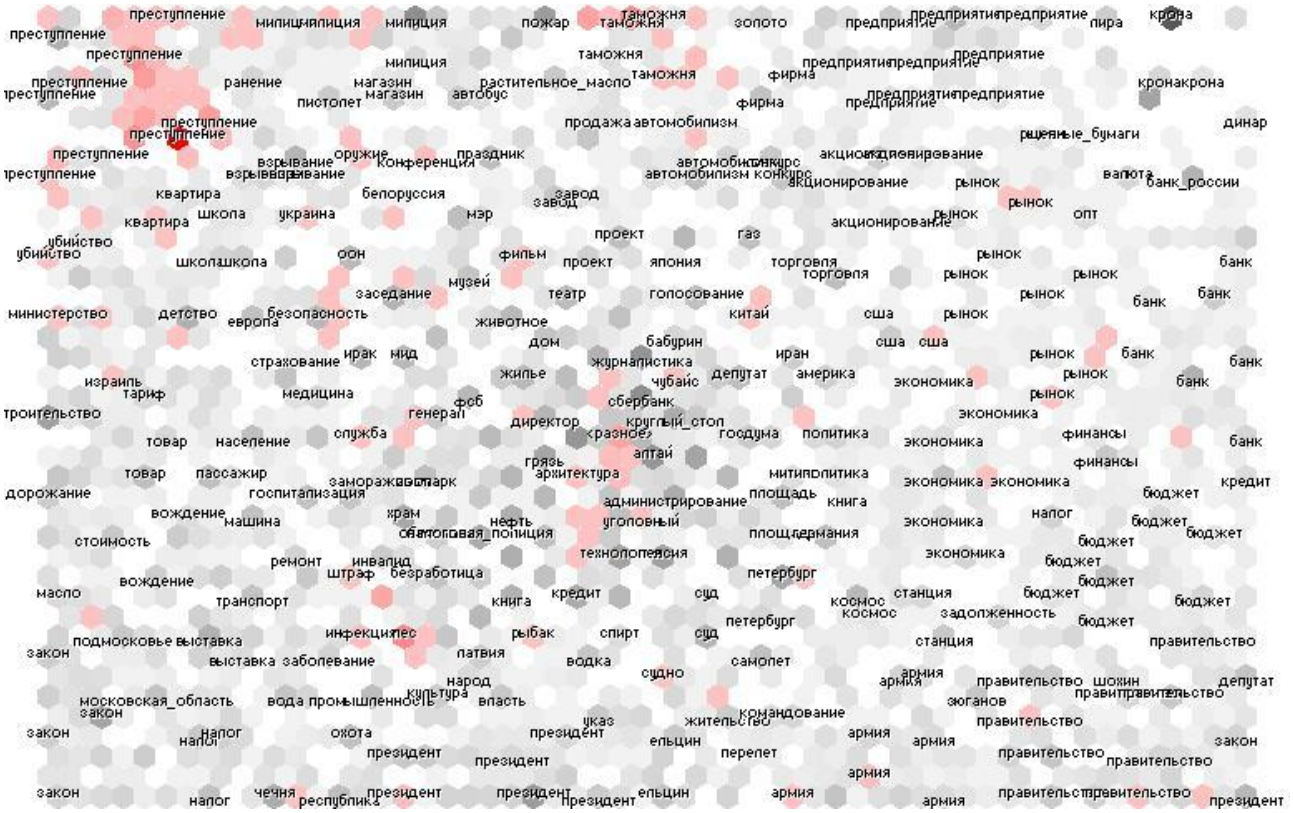

Рис. 2. Тематическая карта TopSOM с подсветкой результатов поиска по запросу "наркотики”.

На рисунке 2 в более мелком масштабе представлена вся карта. На ней красным цветом подсвечены области, которым принадлежат документы, найденные по запросу "*наркотики*".

Как видно, основные скопления документов, содержащих слово "*наркотики*", принадлежат к области, в которых преобладает тема "*преступление*". Еще часть документов относится к тематике, связанной с "*таможней*" и "*инфекцией*", а некоторые документы рассеяны по прочим областям.

Выбрав интересующую область, например в окрестности темы "*таможня*", можно детально просмотреть фрагмент карты в увеличенном масштабе и выбрать документы, связанные, к примеру, с наркокурьерами.

Разработанный комплекс средств поддержки интерфейса пользователя включает возможности масштабирования карты и движения по ней, получения документов из выбранной области, а также контекстного поиска документов по

содержащимся в них словам, с графическим отображением распределения найденных материалов на карте.

Ознакомиться с полной on-line демонстрацией технологии TopSOM для русского и английского языков возможно по адресу: *http://research.metric.ru/*

Заключение

Описанный метод визуализации информационных массивов на основе тематических карт представляет наглядный и удобный для пользователя способ доступа к документам. В дополнение к этому, TopSOM представляет собой полигон для социологических исследований. С его помощью можно, например, отслеживать отслеживать эволюцию тематики и акцентов в потоках поступающих документов за определенные отрезки времени.

Разработанная модификация метода демонстрирует лучшее качество построения карты по сравнению с известными примерами, полученными на базе традиционного варианта метода. На наш взгляд, в настоящей реализации технология тематических карт TopSOM может быть внедрена в действующие поисковые системы.

В заключение отметим, что серьезной проблемой при обработке больших коллекций документов является длительность обучения сети Кохонена в процессе построения карты. Кроме того, при увеличении объема коллекции естественно ухудшается точность аппроксимации локальных свойств семантического пространства документов, которое является многомерным. В случае повышения размера карты, увеличивающего точность классификации, начинает теряться основное достоинство метода – емкость представления информации, дающая возможность окинуть тематику в целом.

Наши эксперименты дают основание полагать, что оптимальный объем визуализируемой коллекции документов ограничен пределом 50-ти тысяч, а размер карты не должен превышать 100x100.

Литература

1. Honkela T., Kaski S., Lagus K., Kohonen T. Newsgroup Exploration with WEBSOM Method and Browsing Interface. // Report A32, Helsinki Univ. of Technology, Laboratory of Computer and Information Science, January 1996.

2. Kohonen T. Analisis of a simple self-organizing process // Biol. Cybern. - 1982. - Vol.44, N.1. - P.135-140.

3. Kohonen T. Self-Organizing Maps. // Springer Series in Information Sciences, Vol. 30, Springer, Berlin, Heidelberg, New York, 1995.

4. Ritter H., Kohonen T. Self-organizing semantic maps. // Biological Cybernetics, Vol.61, P. 241-254, 1989.

5. Honkela T. Comparisons of self-organized word category maps. // In Proc. of Workshop of Self-Organizing Maps 1997 (WSOM'97), Espoo, Finland, June 1997.

6. Харламов А.А., Ермаков А.Е., Кузнецов Д.М. Технология обработки текстовой информации с опорой на семантическое представление на основе

иерархических структур из динамических нейронных сетей, управляемых механизмом внимания // Информационные технологии. - 1998. - N 2. - С. 26-32.

7. Ермаков А.Е. Тематический анализ текста с выявлением сверхфразовой структуры // Информационные технологии. - 2000. - N 11.

TopSOM: visualization of document collections by means of self-organizing maps of topics

Pleshko, V.V., leading developer, "Garant-Park-Internet" Ltd.
Ermakov, A.E., Ph.D., leading developer, "Garant-Park-Internet" Ltd.
Lipinsky, G.B., project manager, "Garant-Park-Internet" Ltd.

Abstarct

A modification of WebSOM method which is widely used for visual representation of full-text document collections was developed. The method proposed is called TopSOM and differs from original WebSOM method in the way semantic document space is formed. Instead of word category map, a new approach is used to extract most meaningfull topics (words and phrases) from documents. TopSOM method generally outperforms WebSOM in quality of clusters and labels generated and does not degrade on small corpora while original method does.