УДК 007:681.518+002.5/.6

Поиск знаний в информационных сетях: базовые модели и технологии (обзор методов и моделей)

В.А. Береговой, Г.П. Крачун, Н.Г. Леонова

Приднестровский государственный университет им. Т.Г.Шевченко, НИИ Медицинских исследований медицинского факультета, НИЛ «ГИППОКРАТ»

Сегодня в информационных хранилищах, распределенных по всему миру, собраны терабайты текстовых данных. Неструктурированные данные составляют большую часть информации, с которой имеют дело пользователи [9]. Эти данные составляют не менее 90% всей информации, а 10% приходится на структурированные данные, загружаемые в реляционные СУБД (системы управления базами данных). Найти в этой информации ценные материалы представляется возможным лишь посредством применения специализированных технологий. Основу развития последних составили исследования и достижения по математическому моделированию [5, 6, 8, 10, 21]. Будучи аппроксимированы в компьютерные технологии, математические модели (исследуемого объекта, предметной области или процесса) приобретают качества и свойства информационной модели, что помогает исследовать особенности изучаемого объекта, а также осуществить разработку изменения картины явления во времени и в пространстве [2, 4, 7, 12, 16, 17, 20]. С другой стороны, компьютерные технологии все шире становятся существенно важным инструментом для математического моделирования [14, 19, 22].

Ряд исследователей использовали возможности моделей реагировать (подобно мозгу и его структурам) на входной поток информации, что позволяет выявить на выходе модели решения (результат) интеллектуальной - по сути - задачи. Так возникла принципиально новая технология “Искусственного интеллекта”, в основе которой – использование знаний (для экспертных целей). Исследования и разработки в рамках направления ”Искусственный интеллект” позволили создать ряд самообучающихся экспертных систем, а также коммерческих направлений применения.

С позиций достижений искусственного интеллекта – знания – метаданные, поскольку они применительно к компьютерной технологии искусственного интеллекта представляют собой структурированный комплекс данных, со своими принципами организации, структурно-функциональными связями и даже законами. В целом же они представляют собой формализованную информацию, позволяющую осуществить целостное описание того или иного объекта с помощью построенной информационной модели, а также получить логически обоснованные выводы [2, 4, 7, 17, 19, 20].

Ядро базы знаний составляют методы, алгоритмы, программы решения различных задач в избранной предметной области. В контексте сказанного следует подчеркнуть, что знания о реальных объектах могут использоваться в компьютерной технологии только тогда, когда осуществлено моделирование знаний. Именно в такой форме они приобретают качества информационной модели и могут быть применены с целью компьютерного решения в конкретной предметной области (независимо от ее отраслевой принадлежности). Одной из форм представления знаний являются сетевые модели (нейросети), которые моделируют в формализованном виде знания, включая учет и отражение в модели множественных отношений между понятиями, а также формализацию представления и применения знаний для логического решения задач в исследуемой предметной области [16, 17, 18, 19] .

Современное информационное общество использует ряд аналитических подходов к представлению информации для обеспечения ее последующего поиска [1]. Один из них -

базируется на теории множеств, другой – на элементах векторной алгебры. Оба подхода эффективно реализовываются в условиях практики, однако у них есть общий недостаток, заключающийся в том, что смысл документа и его базовое содержание в процессе поиска определяется множеством ключевых слов (терминов и понятий). Такие подходы частично ведут к потере содержательных оттенков текста, зато позволяют выполнять быстрый поиск и группировку документов по формальным признакам.

На сегодняшний день разработаны технологии и другие аналитические подходы (например, семантические), в рамках которых делаются попытки выявить смысл текста за счет анализа грамматики текста, использования баз знаний и различных тезаурусов, определяющих семантические связи между отдельными словами и их группами [3]. Такие подходы требуют больших затрат на поддержку баз знаний, тезаурусов для каждого языка, тематики и вида документов.

В современной практике использования информационных технологий нашли применение ряд моделей, позволяющие представить данные большей размерности. К ним следует отнести булевы модели и векторные модели поиска. Последнее обстоятельство придает определенную степень универсальности принципам поиска во многих отраслях знаний – технике, медицине и здравоохранении, общественных науках и др., что позволяет разработать принципы общих подходов в познавательных процессах применительно к формирующемуся информационному обществу.

Булева модель поиска

Булева модель является классической и широко используется с целью представления информации [6]. Она основана на теории множеств, и, следовательно, обладает возможностями информационного поиска. Популярность этой модели обусловлена простотой ее реализации, позволяющей в массивах документов большого объема индексировать и выполнять поиск.

В рамках булевой модели документы и запросы представляются в виде множества ключевых слов - термов. Пусть документальный массив С состоит из множества документов $d_1, d_2, ..., d_n$, а документ d_i, содержит множество различных термов $T(d_i)$. Обозначим через $T = \sum_{i=1}^{n} T(d_i)$ словарь массива С, представляющий собой множество всех термов, встречающихся в документах из С, и через $T(d_i)$ – словарь документа d_i. В булевой модели запрос пользователя представляет собой логическое выражение, в котором ключевые слова (термы запроса) связаны логическими операторами AND, OR и NOT. В различных поисковых системах пользователи могут применять умолчания, не используя в явном виде логических операций, осуществляя перечисление ключевых слов. Чаще всего по умолчанию предполагается, что все ключевые слова соединяются логической операцией AND – в этих случаях в результаты поиска включаются только те документы, которые содержат одновременно все ключевые слова запроса. В тех системах, в которых пробел между словами приравнивается к оператору OR, в результаты поиска включаются документы, в которые входит хотя бы одно из ключевых слов запроса.

При использовании булевой алгебры база данных включает индекс, организуемый в виде инвертированного массива. В нем каждый терм из словаря базы данных содержит список документов, в котором этот терм встречается. В индексе могут храниться также значения частоты вхождения данного терма в каждом документе, что позволяет сортировать список по убыванию частоты вхождения. Классическая база данных, соответствующая булевой модели, организована таким образом, чтобы по каждому терму можно было быстро получить доступ к соответствующему списку документов. Кроме того, структура инвертированного массива обеспечивает его быструю модификацию при включении в базу данных новых документов.

Векторная модель

Большинство известных информационно-поисковых систем и систем классификации информации в той или иной мере основываются на использовании векторной модели описания данных. Векторная модель является классической алгебраической моделью [1]. В рамках этой модели документ описывается вектором в некотором евклидовом пространстве. Каждому используемому в документе терму ставится в соответствие его весовой коэффициент. Описание запроса, соответствующего тематике пользователя представляет собой вектор в том же евклидовом пространстве термов. Для оценки близости запроса и документа используется скалярное произведение соответствующих векторов описания тематики и документа.

В рамках этой модели с каждым термом t_i в документе d_i и запросе q сопоставляется некоторый неотрицательный вес w_{ij}. Таким образом, каждый документ и запрос могут быть представлены в виде k - мерного вектора, $\|w_{ij}\|_{i=1,\dots,k}$, где k общее количество различных термов во всех документах. Согласно векторной модели, близость документа d_j к запросу q оценивается как корреляция между векторами. Эта корреляция может быть вычислена как скалярное произведение соответствующих векторов описаний документов. Для построения векторной модели весовые коэффициенты отдельных термов можно вычислить различными способами. В случае, когда доступна статистика использования во всем информационном массиве данных, более эффективным является применение следующего правила вычисления весов:

$$w_{ij} = tf \times idf_{ij} = tf_{ij} \times \log N / n_i,$$

где n - число документов, в которых используется терм t_i, а N - общее число документов в массиве информационного поиска.

Такой метод «взвешивания» термов может быть обозначен - $tf \times idf$, где tf указывает на частоту использования термина в документе, а idf - на величину, обратную числу документов массива, содержащих данный терм. Значения весов w_{ij} нормируются, т.е. делятся на квадратный корень из суммы весов всех термов, входящих в документ, что позволяет рассматривать документ как ортонормированный вектор.

Когда возникает задача определения тематической близости двух документов или документа и запроса, в модели используется простое скалярное произведение $sim(d_1, d_2)$ двух векторов $\|w_{i1}\|_{i=1,\dots,k}$ и $\|w_{i2}\|_{i=1,\dots,k}$, которое соответствует косинусу угла между векторами-образами документов d_1 и d_2. Очевидно $sim(d_1, d_2)$ принадлежит диапазону [0,1]. Чем больше величина $sim(d_1, d_2)$, тем более близки документы d_1 и d_2. Для любого документа d_i имеем $sim(d_1, d_2)$=1. Аналогично мерой близости запроса q к документу d_i считается величина $sim(q, d_i)$.

На практике чаще всего используются *гибридные модели поиска*, в которых объединены возможности булевой и векторной моделей. В некоторых случаях к ним функционально добавляются некоторые из существующих оригинальных методов семантической обработки информации. Чаще всего в информационно-поисковых системах процедура быстрого поиска выполняется в соответствии с булевой моделью, а результаты ранжируются по весам в соответствии с векторной моделью. Этот подход позволяет качественным образом выявлять информацию в информационном поле при поиске нужного блока данных. Обе модели являются достаточно универсальными применительно к поиску знаний в различных отраслях.

Группировка текстовых данных

Описанные выше модели представления данных имеют общий недостаток, который характеризуется большой размерностью векторного пространства (векторная модель) и множества информационных данных (булева модель). Для обеспечения эффективной работы необходимо сгруппировать как подмножества термов, так и тематически подобные документы. Только в этом случае может быть обеспечена обработка информационных массивов в режиме реального времени. В этом случае используются два основных приема группировки – *классификация* и *кластеризация* [13]. Классификация и кластеризация представляют собой два противоположных полюса применительно к участию пользователя в процессе группировки документов в информационной сети. Осуществление на практике возможности автоматической группировки тематически близких документов позволяет выстроить тематические каталоги. Механизм классификации обычно запускается на отобранных документах только после завершения стадии автоматического выявления сгруппированных данных (кластеров).

Процесс кластеризации - это разбиение множества документов на кластеры, являющиеся, по сути, подмножествами, со смысловыми параметрами, которые заранее известны. Количество кластеров может быть произвольным или фиксированным. Основная идея современных методов кластеризации – это снижение размерности пространства признаков, по которым происходит классификация документов. Задачей кластеризации документов является автоматическое выявление групп подобных документов, исходя из их семантики. Цель всех методов кластеризации заключается в том, чтобы схожесть документов, находящихся в конкретном кластере, были бы максимально близки по своей семантике.

Начальным пространством признаков обычно является пространство термов. Оно обладает способностью сжиматься в информационном поле в процессе анализа большого массива документов. Для его проведения используются различные способы моделирования – весовой, вероятностный, семантический, которые определяют возможности и правила классификации. Актуальной задачей в процессе использования векторной модели в информационно-поисковых системах является задача снижения размерности представленных данных. Этот процесс повышает скорость обработки и выполнения поиска по заданному векторной моделью запроса документов.

Таблица взаимосвязей понятий

В качестве основы для группировки документов в информационном массиве рассматривают семантически значимые понятия. Аналогично случаям отдельных термов, кластеризация документов сопоставляется с кластеризацией понятий, которые более точно отражают тематические свойства документов. Процедура построения таблицы взаимосвязей понятий позволяет практически выявить взаимосвязанные понятия, осуществить их перегруппировку, а также визуализацию поступающего документального массива [9]. Построение таблиц взаимосвязей понятий базируется на языковых средствах информационно-поисковой системы, а также методах кластер-анализа. Семантическое значение понятий определяется на основе информационно-поискового языка.

Таблица взаимосвязей понятий, которая строится в форме статистического анализа (отчета), отражает близость и родственность отдельных понятий из реального мира. Она имеет вид симметричной матрицы $A = \|a_{ij}\|$, элементы которой a_{ij} - коэффициенты взаимосвязи соответствующих пар понятий. Коэффициент a_{ii} отражает количество документов входящего информационного потока. Документы содержат понятия - термины или словосочетания, представленные на языке запросов в соответствии с

понятием i. Со своей стороны, коэффициент a_{ij} - количество документов во входном потоке, которые одновременно соответствуют понятиям i и j. Следует предположить, что качественные признаки документов вполне адекватно выражаются информационно-поисковым языком.

С целью выявления блоков множеств взаимосвязанных понятий в потоке информационного поля применяется алгоритм кластер-анализа. Например, для выделения двух таких блоков необходимо выделить два понятия-полюса (соответствующих, например, индексам k и l), наиболее тесно связанные с другими понятиями (но минимально связанные между собой). Указанные условия можно записать следующим образом:

$$\begin{cases} \sum (a_{ik} - a_{kk}) \to \max, \\ \sum (a_{il} - a_{ll}) \to \max, \\ a_{kl} \to \min, k \neq l \end{cases}$$

Процедура построения таблицы взаимосвязей понятий на первом этапе принимает на своем входе два потока – документальный массив и таблицу понятий, строки которой представляют собой названия понятий и запрос на информационно-поисковом языке, соответствующий этому разделу. На этапе построения таблицы взаимосвязей понятий должен быть создан текстовый файл взаимосвязей понятий, который соответствует матрице $A = \{a_{ij}\}$, где a_{ij} - коэффициенты взаимосвязей понятий i и j. В файле, который должен соответствовать матрице A, первая строка будет определять первое понятие, и она заполняется коэффициентами взаимосвязей с другими понятиями. Коэффициент a_{ij} будет соответствовать количеству документов во входном массиве информации, которые соответствуют понятию i, а коэффициент a_{ij} - количеству документов, которые одновременно определяют понятиям i и j.

На втором этапе построения таблицы взаимосвязей понятий выполняется перегруппировка понятий в зависимости от значений элементов матрицы A. Перегруппировка происходит путем одновременной перестановки строк и столбцов этой матрицы - с целью сведения ее к блочно-диагональному виду. Диагональные блоки соответствуют кластерам обобщенных понятий.

На третьем этапе процедуры происходит визуализация таблицы взаимосвязей понятий для их удобного представления.

На последнем (заключительном) этапе осуществляется формирование типовых запросов для последующей группировки документов, т.е. реализуются механизмы фрагментации документального массива, что облегчает и конкретизирует поиск документов в информационном поле.

Вероятностная модель

Вероятностная модель поиска базируется на теоретических подходах байесовских условных вероятностей [21]. Основным подходом вероятностной модели является вероятностная оценка веса термов в документе. С другой стороны, в качестве оценки соответствия документа запросу используется вероятность того, что пользователь признает документ релевантным.

При описании вероятностной модели используется словарь массива информационных данных, включающий все термы, встречающиеся хотя бы в одном документе из данного массива. С документом сопоставляется вектор $x = (t_1, ..., t_n)$, компонента i которого равна 1, если терм i входит в данный документ, и 0 – в

противном случае. Терм здесь задается своим порядковым номером в словаре, а n - является общим количеством термов в словаре.

Рассмотрим и опишем построение вероятностной модели [7]. Будем считать фиксированным некоторый запрос q, и обозначим через W_1 событие, состоящее в том, что рассматриваемый документ релевантен запросу q; а через W_2 - событие, состоящее в том, что рассматриваемый документ не релевантен запросу q. В этом случае $P(W_i|x)$ - вероятность того, что для документа x наступает событие W_i. Используя эту вероятность, можно сформулировать следующее правило для поиска информации. Так, если $P(W_1|x) > P(W_2|x)$, то документ, представленный вектором x, релевантен запросу q (т.е. он соответствует запросу пользователя). Отметим, что теорема Байеса позволяет перейти к вероятностям, значения которых удобнее оценить следующим равенством:

$$P(W_i|x) = P(x|W_i)\ P(W)/P(x).$$

В приведенной вероятностной модели используется упрощение, заключающееся в предположении о независимости вхождения в документ любой пары термов. В этом случае вероятностная модель приобретает следующий вид:

$$P(x|W_i) = P(x_1|W_1) \times ... \times P(x_n|W_i).$$

С целью упрощения аналитического способа задания модели используем следующие обозначения: $p_i = P(x_i = 1|w_1)$, $q_i = P(x_i = 1|w_2)$. В этом случае получим:

$$P(x|W_1) = \prod_{i=1,...,n} p_i^{x_i} (1-p_i)^{1-x_i},$$

$$P(x|W_2) = \prod_{i=1,...,n} q_i^{x_i} (1-q_i)^{1-x_i}.$$

Неравенство, определяющее релевантность документа запросу, можно представить в следующем виде:

$$\log(P(x|W_1)P(W_1)/P(x|W_2)P(W_2)) > 0.$$

Для ввода весов в представленную модель обозначим через N - общее число документов в информационном массиве; R - число документов, релевантных запросу q; n_i - число документов, в которых имеется терм с номером i; r_i - число документов, релевантных запросу q и включающих терм с номером i. В этих обозначениях $p_i \approx r_i / R$ и $q_i \approx n_i - r_i / N - R$. В качестве веса терма с номером i в документе, представленном вектором x можно взять величину

$$W(i) = \log(r_i \times (\mathrm{N - R - n + r_i})/(n_i - r_i)(R - r_i)).$$

При выполнении информационного поиска, благодаря режиму обратной связи по релевантности, можно итеративным путем уточнять вес термов. В начале поиска вес терма i вычисляется по формуле:

$$W(i) = \log(N - n_i / n_i) \approx \log(N / n_i).$$

Затем на каждом шаге поиска можно определить множество документов, отмеченных пользователем, соответствующих его информационным потребностям. Их общее число можно принять за оценку величины R; число отмеченных документов, содержащих термы с номером i, служит основой оценки величины r_i.

Основы технологии Text Mining как формы поиска данных в информационных потоках

Главная проблема современных коммуникаций – это извлечение действительно ценных сведений из информационных потоков, или, другими словами, получение знаний из информации [9]. Раннее представлялись перспективными системы искусственного

интеллекта, экспертные системы со своими парадигмами фреймов и правил – базы знаний. В конце 80-х годах XX века популярность таких идей иссякла по причинам: не до конца сформировалась общественная потребность в широком использовании таких систем; недостаточными были мощности компьютеров; недоработанными оказались теоретические и алгоритмические основы информационно-поисковых систем. За прошедшее с тех пор время сложилось понимание того, что для решения проблемы информационного хаоса больше подходят технологии, порожденные некогда таким направлением, как контент-анализ, получившее сегодня название Data Mining и Text Mining [3].

Технологии глубинной разработки текста исторически предшествовало создание технологии глубинной добычи данных (Data Mining), методология и подходы которой широко используются и в методах Text Mining [18, 23]. Актуальность технологии Data Mining состоит в том, что она позволяет дополнительно в “сырых данных” выявить практически полезные знания, необходимых для принятия решений в различных сферах трудовой деятельности.

Data Mining – это процесс выделения из данных неявной и неструктурированной информации и представления ее в виде, пригодном для реализации.

Data Mining – это процесс, цель которого – обнаружить новые значимые корреляции, образцы и тенденции в результате просеивания большого объема хранимых данных с использованием методик распознавания образцов плюс применение статистических и математических методов.

Data Mining дословно переводится как «добыча» или «раскопка данных». Однако нередко используют и другие термины: «обнаружение знаний в базах данных» или «интеллектуальный анализ данных», которые можно считать синонимами Data Mining. Идеология Data Mining появилась на стыке прикладной статистики, искусственного интеллекта и баз данных. Фактически рождению нового направления в анализе данных способствовало появление мощных компьютеров и совершенствование технологий записи и хранения данных.

Системы добычи данных в большей степени ориентированы на практическое приложение полученных результатов, чем на выяснение природы исследуемого явления. При использовании технологии Data Mining исследователя не очень интересует конкретный вид зависимостей между переменными решаемой задачи. Основное внимание уделяется поиску решений, на основе которых можно было бы строить достоверные прогнозы. В процедуре добычи данных преобладает содержательный подход к изучению знаний, а критерием качества применяемых методов является их практическая реализация.

Новое направление в обработке текстовой информации – “глубинная разработка текстов” (Text Mining) – это алгоритмическое выявление прежде неизвестных связей и корреляций в уже имеющихся текстовых данных. Задача Text Mining – отобрать ключевую и наиболее значимую информацию для пользователя. Разработанные на основе статистического и лингвистического анализа, а также методов искусственного интеллекта, технология Text Mining предназначена для проведения смыслового анализа, обеспечения навигации и поиска в неструктурированных текстах. Кроме того, Text Mining в отличие от традиционных подходов, не только находит списки документов, релевантных запросам, но обеспечивает и новый уровень семантического поиска документов.

Оформившись в середине 90-х годов XX века как направление анализа неструктурированных текстов, технология Text Mining сразу же взяла на вооружение методы классической добычи данных, например, таких как классификация или кластеризация. В Text Mining появились и дополнительные возможности - автоматическое реферирование текстов и выявление феноменов, т.е. понятий и фактов. Важный компонент технологии Text Mining - извлечение из текста характерных элементов, свойств, которые затем могут использоваться в процессе поиска в качестве метаданных документа, ключевых слов, аннотаций. Другая важная задача состоит в отнесении

документа к некоторым категориям из заданной схемы систематизации. Следует подчеркнуть, что технологии Text Mining присуща абсолютная объективность – в ней отсутствует субъективизм, свойственный человеку-аналитику.

Технологии Data Mining и Text Mining являются новой тенденцией в развитии средств и методов обработки данных. Они помогают найти скрытые закономерности и отношения в исследуемых данных для того, чтобы можно было бы принять более обоснованные решения. Сфера их применения ничем не ограничена – они везде, где имеются какие-либо данные. Но в первую очередь в применении новых технологий сегодня наиболее заинтересованы коммерческие предприятия, проектирующие и внедряющие проекты на основе информационных хранилищ данных. Новые технологии представляют большую ценность для руководителей и аналитиков в их повседневной деятельности. Опыт многих таких предприятий показывает, что отдача от использования технологий Data Mining и Text Mining может достигать 1000% [18].

Основные элементы Text Mining

В соответствии с уже сложившейся концепцией [3], к основным элементам Text Mining относятся реферирование, выявление феноменов, классификация, кластеризация, ответ на вопросы, тематическое индексирование, поиск по ключевым словам, средства поддержки и создания тезаурусов.

Обычно выделяют четыре основных вида приложений технологии Text Mining.

1. Классификация текста, в которой используются статистические корреляции для построения правил размещения документов в предопределенные категории.

2. Кластеризация, базирующаяся на признаках документов. Используются лингвистические и математические методы без применения предопределенных категорий.

3. Построение семантической сети или анализ связей, которые определяют появление дескрипторов (ключевых фраз) в документе для обеспечения поиска и навигации.

4. Извлечение фактов из текста с целью улучшения классификации, поиска и кластеризации.

Чаще всего решаемая в Text Mining задача – это классификация, т.е. отнесение объектов базы данных к заранее определенным категориям. Фактически задача классификации – это вариант классической задачи распознавания, когда система по обучающей выборке относит новый объект к той или иной категории. Особенность же системы Text Mining состоит лишь в том, что количество таких объектов и их атрибутов может быть очень большим. Поэтому должны быть предусмотрены интеллектуальные механизмы оптимизации процесса классификации. В существующих сегодня системах классификация применяется, например, для решения таких задач, как группировка документов, размещение документов в определенные папки, сортировка сообщений электронной почты, избирательное распространение новостей подписчикам и др.

Другая задача, основанная на кластеризации, состоит в выделении компактных подгрупп объектов с близкими свойствами. Система должна самостоятельно найти признаки и разделить объекты по подгруппам. Решение этой задачи, как правило, предшествует задаче классификации, поскольку позволяет определить группы объектов. В процессе кластеризации строится базис ссылок от документа к документу, основанный на весах и совместном употреблении определяемых ключевых слов. Сегодня кластеризация широко применяется при реферировании больших документальных массивов или определении взаимосвязанных групп документов, а также для упрощения процесса просмотра при поиске необходимой информации, для нахождения уникальных документов из коллекции, для выявления дубликатов или очень близких по содержанию документов.

К числу задач, которые можно решать средствами технологии Text Mining, можно отнести, например, прогнозирование, которое состоит в том, чтобы предсказать по значениям одних признаков - значения остальных. Еще одна задача – нахождение исключений, т.е. поиск объектов, которые своими характеристиками сильно выделяются из общей массы. Для этого сначала выясняются средние параметры объектов, а затем исследуются те объекты, параметры которых наиболее сильно отличаются от средних значений. Как правило, поиск исключений проводится после классификации или кластеризации - для того чтобы выяснить, насколько были точны последние.

Несколько отдельно от кластеризации стоит задача поиска связанных признаков (полей, понятий) отдельных документов. От прогнозирования эта задача отличается тем, что заранее неизвестно, по каким именно признакам реализуется взаимосвязь. Цель именно в том и состоит, чтобы найти связи между отдельными признаками. Эта задача сходна с кластеризацией, но выполняется не по множеству документов, а по множеству присущих документу признаков.

И, наконец, для обработки и интерпретации результатов Text Mining большое значение приобретает визуализация данных, что подразумевает обработку структурированных числовых данных. Однако визуализация также является ключевым звеном при представлении данных неструктурированных текстовых документов. В частности, современные системы класса Text Mining могут осуществлять анализ больших массивов документов и формировать предметные указатели понятий и тем, освещенных в этих документах. Визуализация обычно используется как средство представления контента всего массива документов, а также для реализации навигационного механизма, который может применяться при исследовании документов и их классов.

Заключение

Таким образом, вышеизложенные материалы отражают всю совокупность сложных технологий поиска данных в информационных потоках. Существующие технологии построены на моделировании характера и сути процессов, происходящих внутри информационного потока. Они отражают достижения в области теории вероятности, математической статистики, теории распознавания образов и методов искусственного интеллекта. С другой стороны, современные компьютерные технологии уже сегодня имеют в своем распоряжении ряд новых подходов, которые в перспективе с большой степенью точности позволят выявлять данные и знания в глобальных информационных потоках, что в будущем обеспечит не только создание новых компьютерных технологий, но и позволит открыть совершенно новые законы в познании окружающей человека природы и мира в целом.

Литература

1. Аникин В.М. Аналитические модели детерминированного хаоса. М.: Физматлит, 2007.- 328 с.
2. Артемьев В.И. Обзор способов и средств построения информационных приложений//Системы управления базами данных. - 1998. - №3. - с. 71-80.
3. Барсегян А.А, Куприянов М.С., Степаненко В.В. Холод И.И. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, Olap/2-е изд., перераб. СД-СПб:BHV- Санкт – Петербург, 2008.-384 с.
4. Брукшир Дж. Введение в компьютерные науки.- М.: Диалог - МИФИ, 2001.- 688 с.
5. Васильков Ю.В., Васильева Н.Н. Компьютерные технологии вычислений в математическом моделировании.- М.: Финансы и статистика, 2002.- 256 с.

6. Горбатов В.А. Фундаментальные основы дискретной математики. Информационная математика. М.: Высш. шк., 2000.- 544 с.
7. Городецкий А.Я. Информационные системы. Вероятностные модели и статистические решения.- СПб: Изд-во СПб ГПУ, 2003. - 326 с.
8. Грэхем Р., Кнут Д., Паташник О. Конкретная математика: Основание информатики.- М.: Мит, 1998.- 703 с.
9. Гусев В.С. Поиск в Internet.- СПб.: Диалектика, 2004.- 336 с.
10. Дьяконов В.Г. Компьютерная математика: Теория и практика.- М.:Нолидж, 2001.- 1295 с.
11. Жамбю М. Иерархический кластер-анализ и соответствие/ Пер. с франц. Б.Г. Миркина. – М.: Финансы и статистика, 1988. – 342 с.
12. Йордан Эд., Аргила К. Объектно-ориентированный анализ и проектирование систем. -М.: Изд-во “Лори”,2006.-288 с.
13. Кириченко К.М., Герасимов М.Б. Обзор методов кластеризации текстовой информации.//Дост. из URZ:http://www.dialog-21.ru/Archive/2001/volume2/2-26.htm(24 июня 2009) /Материалы межд. конфер.”Диалог-2001”.
14. Кошкарев А.В., Тикунов В.С. Геоинформатика. – М.: Картегеоцентр - Геодезиздат, 1993.-213 с
15. Мандель И.Д. Кластерный анализ/ Предисл. Б.Г.Миркина.– М.: Финансы и статистика, 1988. – 176 с.
16. Олифер В.Г., Олифер Н.А. Компьютерные сети: Принципы, технологии, протоколы.: Учебник. - СПб.: Питер, 2001.- 672 с.
17. Пуха Ю. Объектные технологии построения распределенных информационных систем // Системы управления базами данных. - 1997. - №3. - с. 4-20.
18. Самойленко А.П., Дюк В.А. Data Mining: Учебный курс.CD.СПб: Питер.: 2001.- 368 с.
19. Сахаров А.А. Концепции построения и реализации информационных систем, ориентированных на анализ данных// Системы управления базами данных. - 1996. - №4. - с. 5-70.
20. Советов Б.Я. Информационные технологии: Учеб. для вузов. – 4-изд., стер. -М.: Высш. шк., 2008. - 263 с.
21. Терехов С.А.. Введение в байесовы сети //Школа-семинар “Совр. пробл. нейроинформатики”, 29-31 января 2003. МИФИ, Москва.-V Всеросс. конф. “Нейроинформатика-2003”/Отв.ред. Ю.В. Тюменцев- Часть I: Лекции по нейроинформатике. -М.: МИФИ, 2003.- 188 с. (149-186).
22. Цветков В.Я. Геоинформационные системы и технологии.– М.: Финансы и статистика, 1998.- 288 с.
23. Чубукова И.А. Data Mining: учебное пособие. М.: Интернет-университет информационных технологий - ИНТУИТ.ру, 2006.- 384 с.

Специалист НИЛ «ГИППОКРАТ» В.А. Береговой
НИИ Медицинских исследований
медицинского факультета
ПГУ им.Т.Г. Шевченко

Зав. НИЛ «ГИППОКРАТ», Г.П. Крачун
доцент

Ст. н.с. НИЛ «ГИППОКРАТ», Н.Г Леонова
доцент