электронный научно-технический журнал
ИНЖЕНЕРНЫЙ ВЕСТНИК
издатель ФГБОУ ВПО «Московский государственный технический университет им. Н.Э. Баумана»

Интеллектуальный анализ данных в электронных обучающих системах

77-48211/656610

12, декабрь 2013

Белоус В. В., Домников А. С.

УДК 378.146

Россия, МГТУ им. Н.Э. Баумана
walentina.belous@gmail.com
asdomnikoff@mail.ru

Интеллектуальный анализ данных в электронных обучающих системах

В настоящее время наблюдается бурное развитие электронных образовательных технологий, основанных на возможностях современных персональных вычислительных устройств и ресурсах глобальной сети Интернет. Эта методология генерации и приобретения знаний в англоязычной литературе получила название e-learning, что на русский язык чаще всего переводится как электронное или дистанционное образование.

По сравнению с традиционными образовательными технологиями электронное образование обладает множеством очевидных преимуществ, например, гибкостью, доступностью, мобильностью, экономичность и др. Это стало причиной появления прогнозов о скорой и неизбежной кончине традиционного обучения с учителем и повсеместном распространении средств электронного образования. Опыт эксплуатации многочисленных систем e-learning (Moodle, Sakai, Blackboard и др.) выявил несколько существенных и трудноустранимых недостатков, влияющих на

эффективность электронного образования [5]. По всей видимости, основным дефектом данной методологии является слабая связь между обучаемым и преподавателем или администратором курса. В традиционных образовательных учреждениях очень большое значение имеет личное общение ученика и преподавателя. Оно позволяет получить целостное представление об успехах и проблемах обучаемого и наметить эффективный маршрут продвижения по курсам и модулям. В системах электронного обучения пока не удалось найти полноценной замены этому фактору, что часто приводит к дезориентации обучаемого в образовательном пространстве.

Вместе с тем, любая образовательная активность обучаемого в среде современной системы e-learning отслеживается и фиксируется в многочисленных базах, лог-файлах, персональных профилях и др. Средства электронной обучающей платформы аккумулируют громадные массивы разнородных данных, которые потенциально способны описать текущую ситуацию и перспективы обучаемого. Проблема заключается в извлечении, обработке и структурировании этих данных.

Методы извлечения скрытых данных обсуждаются в сравнительно новой научной дисциплине Data Mining, возникшей на стыке искусственного интеллекта, анализа данных, автоматической классификации и статистических методов обработки информации [5, 15]. На русский язык это название переводят обычно как интеллектуальный анализ данных. Методы интеллектуального анализа данных нашли широкое распространение в управлении бизнес-объектами, принятии решений, проектировании, стратегическом планировании и многих других отраслях человеческой деятельности. Их интеграция в системы электронного и дистанционного обучения началась сравнительно недавно в рамках концепции educational data mining (EDM) [5].

В самом широком смысле целями EDM являются синтез методов и средств, предназначенных для понимания и предсказания образовательных ситуаций, а также разработка инструментов для проектирования образовательных артефактов (модулей, курсов, программ и пр.). В [3] приведен обстоятельный обзор множества публикаций по EDM и выявлены основные направления исследований:

- анализ и визуализация данных;
- синтез обратной связи между студентом и инструктором;
- моделирования поведения студента в образовательных ситуациях;
- прогнозирование и выработка рекомендаций по обучению;
- классификация и кластеризация данных;
- генерация ассоциативных правил;
- конструирование курсов;
- планирование и оперативное управление образовательным процессом.

Кластеризация в системах EDM применяются обычно для разбиения студентов на группы, которые характеризуются близкими значениями некоторых числовых или качественных показателей. Например, в группу могут быть включены студенты по сходству образовательных программ, квалификации, общности целей или интересов, сетевой активности и пр. Задача кластеризации сводится к поиску сгущений точек в пространстве признаков или разрезанию графа, представляющего отношение сходства или подобия между объектами. Для этого используются хорошо известные методы кластеризации, разработанные в прикладной статистике, кластерном анализе и вычислительной математике: статистические алгоритмы, EM-алгоритм, алгоритм K-средних, графовые алгоритмы, алгоритмы семейства FOREL, иерархические алгоритмы, нейронные сети Кохонена, ансамбли

кластеризаторов, алгоритмы семейства KRAB и многое другое [7, 9, 10, 12, 17, 21-26].

Ассоциативные правила (association rules) применяются для формализации шаблонов поведения студента в электронной образовательной среде. С их помощью создаются типовые маршруты обучения и курсовые структуры, ориентированные на целевые аудиторию или отдельных потребителей образовательных услуг. Для решения задач такого применяется аппарат нечеткой математики и так называемый apriori-алгоритм [2, 16].

Множество образовательных ресурсов хранится в текстовой форме. Это могут быть HTML-документы, книги, статьи, доклады, профайлы и многое другое. Методы извлечения информации из текстовых репозиториев рассматриваются в разделе интеллектуального анализа данных, который называется анализ текста (text mining) [5].

Текстовые данные отличаются большим разнообразием не только по семантике, но и по форме представления. Например, текстовым источником может быть структурированный фрагмент базы данных, размеченный документ, записанный в формате HTML или XML, носитель, форматированный по правилам RTF или MIF и т.д. Это требует глубокой предварительной обработки источника, которую, в общем случае, способен выполнить только человек.

В современных системах электронного обучения методы анализа текстов чаще всего применяются для рациональной организации текстовых хранилищ, когда документы, связанные одной темой или близкие по смыслу, объединяются в отдельные кластеры. Текстовый анализ профайлов студентов позволяет сформировать так называемые группы по интересам. Однако потенциальные возможности интеллектуального анализа текстовых выходят далеко за пределы данных направлений. Его можно использовать для автоматического формирования основного контента электронного курса и вспомогательных

учебных материалов. Для этого подсистема анализа текстовых данных анализирует текстовые носители, находит релевантные фрагменты, извлекает их из файлов и формирует семантически однородные таксоны.

Типовая процедура обработки текстовых документов показана на рис. 1.

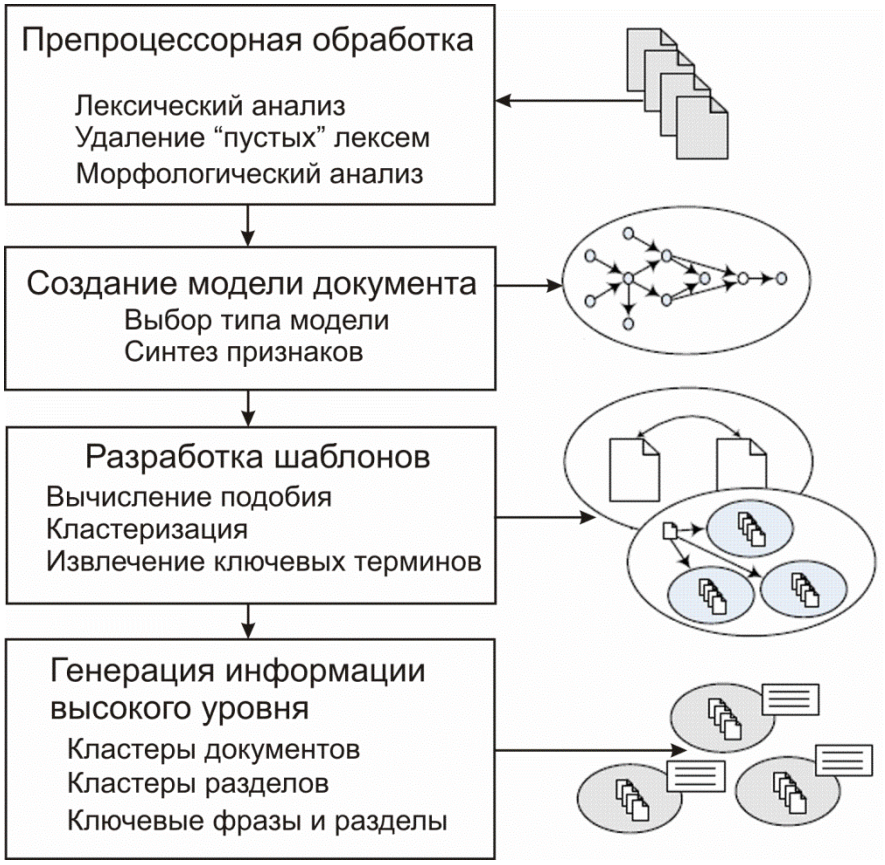

Рис. 1. Типовая процедура обработки документов в text mining

Блок лексического анализа разбивает текстовый массив на отдельные лексемы: основные и служебные. Пустыми называются лексемы, которые не несут содержательной предметной информации. Таковыми могут быть дескрипторы языков разметки, форматирующие операторы форматов RTF или MIF, междометия, артикли и др.

Блок морфологического анализа ищет лексемы, представляющие собой разные формы выражения одних сущностей. Например, слова, которые различаются только падежными окончаниями, или просто являются синонимами.

После реализации всех процедур препроцессорной обработки формируется модель документа, предназначенная для автоматической компьютерной обработки. В настоящее время наибольшее распространение получили две модели:

- представление документа в виде вектора весов в выбранном терминологическом пространстве [18];
- представление документа в виде индексного графа [14].

Графовая или векторная формализация документов позволяет ввести меру близости на множестве текстовых данных. Мера близости – это алгоритмический способ (функция или соответствие), который позволяет оценить подобие любых двух документов при помощи числа или набора чисел. Кластер представляет собой набор документов, меры близости расположенных на небольшом расстоянии друг от друга согласно выбранной метрике.

Обычно кластеры описываются при помощи матриц инцидентности, в которых строки соответствуют документам, столбцы – кластерам, а на пересечении строки и столбца стоит единица тогда и только тогда, когда документ является членом кластера. Этот способ описания предназначен для компьютерной обработки, но не удобен для восприятия человеком, особенно если размеры матрицы велики. Чтобы избавить пользователя от необходимости чтения документов, с каждой клеткой матрицы связывается некоторая ключевая фраза, по содержанию которой можно судить о ее носителе.

Векторная модель документа

Содержательный анализ текстовых документов – это задача очень высокой сложности, которая недоступна современным вычислительным устройствам, несмотря на очевидные успехи в области искусственного интеллекта и интеллектуального анализа данных. Для этого требуется разработать формальное описание документа, пригодное компьютерного анализа и обработки.

Самой распространенной моделью документа является его представление в виде вектора в терминологическом пространстве [5]. Термином будем называть любое одиночное слово (лексему), которое присутствует в документе, прошедшем все процедуры препроцессорной обработки.

Обозначим множество терминов некоторой коллекции документов через $T = \{t_i\}, i = \overline{1,|T|}$. Документы коллекции представляются векторами действительных чисел вида $d = (w_1, w_2, ..., w_n)$, где $n=|T|$, а каждое w_i – есть вес термина в данном документе. Вес представляет собой вычисляемое значение.

В современных исследованиях по анализу текстовых данных эти веса рассчитываются по-разному. В простейшем случае w_i – это просто частота термина в данном документе (совокупное количество вхождений). Часто это число обозначают $w_i=tf_i$ (от английского term frequency).

Более популярной является схема, учитывающая вхождение термина в различные документы. В этом случае вес рассчитывается по формуле $w_i = tf_i \times \log(N / df_i)$, где df_i – количество документов, в которых встретился данный термин (документная частота, document frequency), а N – общее число документов. Данный способ вычисления весов иногда называется TF×IDF (частота термина × обратная документная частота, term frequency × inverse document frequency).

Этот способ измерения весов обладает двумя полезными свойствами. Во-первых, он дает высокие значения для терминов, которые встречаются в небольшом числе документов, тем самым подчеркивая различия данных носителей. Во-вторых, если термин редко встречается в данном документе или входит во множество других носителей, то его информативность будет низкой, и, соответственно, измеритель TF×IDF присвоит ему небольшой вес.

Классификация и кластеризация документов, представленных в виде числовых векторов, основана на предположении, что тематически подобные

документы располагаются на небольшом расстоянии друг от друга терминологическом многомерном пространстве. Вычислительная простота векторной модели послужила причиной ее широкого распространения в современных исследования по интеллектуальному анализу текстов, автоматической классификации и компьютерной лингвистике.

Для принятия объективного решения о близости или несхожести документов, представленных точками в многомерном терминологическом пространстве необходимо предложить некоторые вычисляемые показатели – меры близости. Любая такая мера должна обладать двумя очевидными свойствами:

- давать большие значения для сходных документов;
- различать несходные документы, присваивая им низкие значения на шкале подобия (наоборот, когда подобие измеряется по расстоянию меду документами).

В многочисленных работах по кластерному анализу было предложено значительное число способов измерения близости документов в многомерном пространстве [1, 3, 6, 8, 16]. Все они могут быть разделены на три типа:

- меры, основанные на расстояниях;
- угловые меры;
- корреляционные меры.

Меры, основанные на расстояниях

Если два документа d_i и d_j принадлежат одному кластеру, то расстояние между соответствующими точками в многомерном векторном пространстве должны быть небольшим. Это расстояние будет велико между любыми двумя точками различных кластеров.

Измерение расстояния выполняется при помощи функции, которая называется метрикой или функцией расстояния. Метрика – эта функция

вида $\rho : \mathrm{E}^n \to \mathrm{R}$, где E^n – n-мерное евклидово пространство, а R – множество действительных чисел. Любая функция расстояния должна удовлетворять следующим условиям:

1. $\rho(x, y) \geq 0$ и $\rho(x, y) = 0$ тогда и только тогда, когда $x=y$;
2. $\rho(x, y) = \rho(y, x)$ (симметрия);
3. $\rho(x, y) + \rho(y, \mathrm{z}) \geq \rho(x, z)$ (неравенство треугольника).

В многочисленных публикациях по кластерному анализу и анализу данных предложено несколько десятков различных метрик, большинство из которых являются геометрическими расстояниями в многомерном пространстве и вычисляются по следующей общей формуле: $\rho(\mathrm{X}, \mathrm{Y}) = (\sum_{i=1}^{n} | x_i - y_i |^p)^{1/p}$, где $X=(x_1,\ldots x_n)$, $Y=(y_1,\ldots y_n)$ [21-24]. Чаще всего используются следующие функции расстояния:

- Евклидово расстояние $\rho(\mathrm{X}, \mathrm{Y}) = (\sum_{i=1}^{n} | x_i - y_i |^2)^{1/2}$;
- расстояние Хемминга $\rho(\mathrm{X}, \mathrm{Y}) = \sum_{i=1}^{n} | x_i - y_i |$. Обычно для оценки объектов в пространстве бинарных признаков;
- расстояние Чебышева $\rho(\mathrm{X}, \mathrm{Y}) = \sup \left\{ \left| \mathrm{x_i} - \mathrm{x_i} \right| \right\}$.

Несмотря на интуитивную ясность кластеризации на основе метрик, практика показала, что этот способ измерения близости документов дает не очень надежные результаты.

Меры сходства

Функции расстояния – это далеко не единственный способ измерения подобия между объектами. Широкое применение нашли неметрические способы оценки сходства, которые основаны на иной аксиоматике.

Неотрицательная вещественная функция $\alpha(x, y)$ называется мерой сходства, если выполняются следующие аксиомы:

1. $0 \le \sigma(x, y) < 1, \forall x \ne y$ (неотрицательность);
2. $\sigma(x, x) = 1, \forall x$ (тождественность);
3. $\sigma(x, y) = \sigma(y, x)$ (симметричность).

Чем выше значение меры сходства, тем более похожими будут измеряемые объекты. В это смысле меры сходства коренным образом отличаются от метрик, который считают подобные объекты расположенными близко друг от друга.

Угловые меры

Применение угловых мер основано на предположении, что смысловая близость документов, представленных в виде числовых векторов в многомерном пространстве терминов, задается общим направлением, а не расстоянием между ними. Наибольшее распространение в работах по анализу данных получила косинусная мера.

Пусть $X=(x_1,\ldots x_n)$, $Y=(y_1,\ldots y_n)$ – два вектора в n-мерном пространстве. Скалярное произведение двух векторов рассчитывается по известной формуле $X \bullet Y = x_1 y_1 + x_2 y_2 + \ldots + x_n y_n$. Длина вектора X равна $\| X \| = (\sum_{i=1}^{n} x_i^2)^{1/2}$. Если α – угол между векторами, то косинус этого угла равен

$$\cos\alpha = \frac{X \bullet Y}{\| X \| \times \| Y \|}.$$

Этот способ расчета используется в общей ситуации, когда координаты векторов могут быть отрицательными, а косинус угла лежит в пределах от -1 до 1. Чтобы косинусную меру привести в соответствие с аксиоматикой сходства, используют нормированное значение, вычисляемое по формуле $\cos^*\alpha = 1/2(\cos\alpha + 1)$.

Косинусная мера слабо зависит от расстояния между векторами, в то время как любая метрика основывается на этой характеристике, но не учитывает направление векторов в пространстве.

Корреляционные меры

Кроме разнообразных метрик и угловых мер, имеющих очевидное геометрическое обоснование, в анализе текстовых данных (и вообще, в кластерном анализе) широко применяются способы оценки близости, заимствованные из математической статистики. Это многочисленные корреляционные коэффициенты и меры, оценивающие статистическую значимость зависимости между случайными величинами: скалярными или векторными. Методы корреляционного анализа очень активно применяются в исследованиях по экономике, социологии, биологии, психологии и технических науках. Поэтому не представляется возможным дать сколь-нибудь полное описание способов измерения близости корреляционного типа. Рассмотрим две простые меры, адекватность которых в задачах кластеризации документов подтверждена экспериментально.

Мера Жаккара – это способ измерения сходства, который первоначально предложен для оценки подобия популяций и ландшафтов. Опыт показал, что его универсальность выходит далеко за пределы исследований по географии и биологии.

Мера Жаккара основана на оценке соотношения общих признаков двух объектов к их совокупному количеству. Более точно, пусть имеются два вектора одинаковой длины $X=(x_1,\ldots x_n)$, $Y=(y_1,\ldots y_n)$. Тогда степень их сходства оценивается по формуле $K(\mathrm{X},\mathrm{Y}) = \dfrac{X \bullet Y}{X^2 + Y^2 - X \bullet Y}$.

Можно показать, что при небольших значениях ($K \to 0$) мера Жаккара ведет себя подобно косинусной мере, при $K \to 1$ приобретает свойства, близкие к евклидовой метрике.

Коэффициент корреляции Пирсона (линейный коэффициент корреляции) задается выражением:

$$R(X,Y)=\frac{(X-\overline{X})\bullet(Y-\overline{Y})}{\|X-\overline{X}\|\times\|Y-\overline{Y}\|}, \text{ где } \overline{X}=\frac{1}{n}\sum_{i=1}^{n}x_i, \overline{Y}=\frac{1}{n}\sum_{i=1}^{n}y_i.$$

Значения коэффициента, вычисленного по данной формуле, могут принимать значения в диапазоне от -1 до1. Чтобы использовать его как меру сходства векторов в многомерном пространстве коэффициент нормируют следующим образом: $R^*(X,Y)=\frac{1}{2}\left(\frac{(X-\overline{X})\bullet(Y-\overline{Y})}{\|X-\overline{X}\|\times\|Y-\overline{Y}\|}+1\right)$.

Векторная модель и все ранее рассмотренные меры близости не учитывали действительное семантическое значение терминов для документов набора. Понятно, что в любом документе могут существовать ключевые термины, точно выражающие семантическую принадлежность источника, и случайные, малоинформативные термины, которые не являются характерными для данного дискурса. Если принять во внимание различающую силу выбранных терминов и выбрать способы измерения близости, которые обладают высокой чувствительностью к этой характеристике, то точность кластерного анализа текстовых документов может быть значительно улучшена.

Различающая способность термина задается силой его влияния на среднюю близость документов в наборе. В [19] предложен способ измерения различающей способности термина как выражения:

$$dis(k)=\frac{1}{N}\sum_{i=1}^{N}sim(d_{ik},c_k)-\frac{1}{N}\sum_{i=1}^{N}sim(d_i,c).$$

Здесь: *sim* – это некоторая функция близости между документами, d_{ik} – представление документа *i*-го документа без учета термина *k*, c_k – средний центральный вектор набора документов, подсчитанный без учета термина *k*, *d* и *c* документ и средний вектор, в которых учитывается термин *k*.

Эксперименты с учетом различных мер близости показали, что существуют термины с положительным, отрицательным и нулевым значением меры *dis()*. Термины с положительным значением меры различающей способности делают документы набора менее близкими друг другу (в среднем). Использование терминов с отрицательной мерой уменьшает степень различия, то есть в среднем приближает документы. Термины с нулевой мерой не оказывают влияния на сходство документов.

Исследование различных способов измерения близости с учетов различающей способности терминов показало следующее:

- При работе с метриками термины получают положительное значение различающей способности. Иными словами, удаление термина приводит к уменьшению расстояния между векторами (сходство повышается). И наоборот, введение нового термина во все векторы увеличивает расстояние между ними.
- Угловые меры демонстрируют высокую чувствительность к изменениям терминологического пространства. В частности, мера косинуса способна изменить свое значение в любую сторону, при пополнении пространства новым термином.
- Корреляционные меры такие, как мера Жаккара и коэффициент корреляции Пирсона, ведут себя подобно угловым мерам в задачах кластеризации в многомерных терминологических пространствах. Мера косинуса обладает большей чувствительностью к различающим способностям терминов, чем корреляционные меры.
- Косинусная мера не зависит от длин векторов. Она способна выражать не только различие, но и сходство в употреблении терминов разными документами.

Основным недостатком представления документа в многомерном пространстве является нечувствительность этой модели к смысловым и

лингвистическим зависимостям, которые могут связывать различные слова в документе.

Графовая модель документа

Для описания структурных характеристик документа и взаимосвязи его частей были предложены многочисленные графовые и сетевые модели текста, например семантические сети, клаузальные графы и др. К числу таких моделей относится индексный граф документа (document index graph, DIG), получивший наибольшее распространение в работах по интеллектуальному анализу текстовых данных [5,6,9].

Индексный граф представляет собой ориентированный граф $G=(V,E)$, в котором $V=\{v_1,v_2,\ldots,v_n\}$ – множество вершин, а $E=\{e_1,e_2,\ldots,e_m\}$ – множество дуг. Вершины описывают слова заданной совокупности документов, а дуги представляет упорядоченность слов. Более точно, дуга $e=(v_i,v_j)$ с началом в вершине v_i и окончанием в вершине v_j, существует тогда и только тогда, когда слово v_i предшествует слову v_j в каком-либо из документов заданного набора.

Пусть в некотором документе присутствует предложение, представляющее собой последовательность слов вида $(v_1,v_2,\ldots,v_m)$, тогда в индексном графе существует ориентированный путь (v_1,v_2) $(v_2,v_3)\ldots(v_{m-1},v_m)$ с началом в v_1 *и* окончанием *в* v_m.

Рассмотрим индексный граф на простом наборе из двух документов. В первом документе содержатся два предложения: `HTML – язык разметки документов, HTML-документ`. Второй документ состоит из предложений: `английский язык, текстовый документ, документ на английском языке, язык текста – английский, английский текст`. Индексный граф этого набора показан на рис. 2.

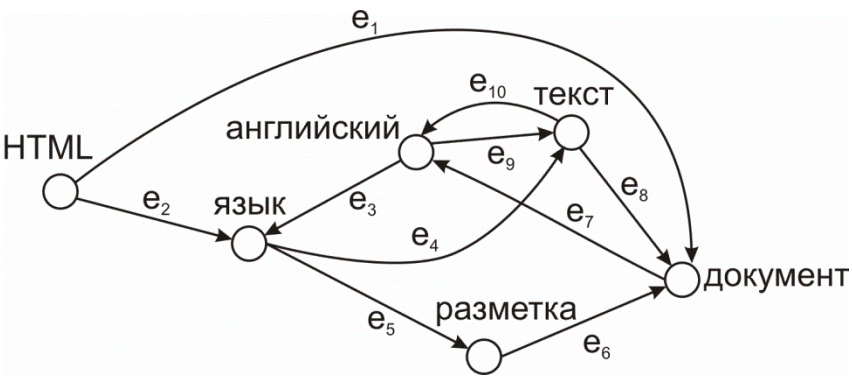

Рис. 2. Пример индексного графа

Пусть задан индексный граф G, описывающий некоторую совокупность документов $D = \{d_i\}$. Тогда любому документу d_i из этого набора соответствует подграф G_i индексного графа G. Для любой пары d_i и d_j документов набора $d_i, d_i \in D$ можно определить степень их соответствия (подобия) M_{ij} по следующему простому правилу $M_{ij} = G_i \bigcap G_j$. В отличие от методик, использующих представление документа в терминологическом векторном пространстве, индексный граф дает более точное измерение сходства документов, поскольку учитывает не только состав терминов, но и их упорядоченность.

В процессе программирования используется более развитая модель, которая вместе с информацией о вершинах и дугах индексного графа хранит дополнительные данные о текстовых источниках. Пример такой модели показан на рис. 3.

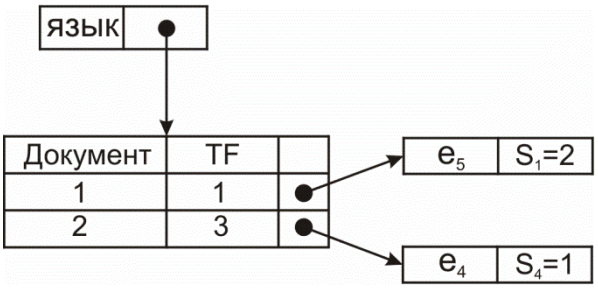

Рис. 3. Структура данных вершины индексного графа

С каждой вершиной графа связана так называемая таблица документов. В ней перечислены номера документов набора, в которых встречается данный термин. Во втором столбце таблицы записаны значения частоты данного термина (term frequency, TF) в документе. В примере, показанном на рис.3, слово «язык» один раз встречается в первом документе (TF=1) и три раза – во втором (TF=2). Последняя ячейка – это ссылка на таблицу дуг, которая хранит данные о последовательности слов в предложениях.

Таблица дуг описывает дуги, исходящие из данной вершины в подграфе индексного графа, соответствующем текстовому носителю. Число строк таблицы равно количеству исходящих дуг в данном подграфе. Первая ячейка строки – это имя дуги, вторая ячейка хранит сведения о позициях данного слова в предложениях документа. Так, в индексном подграфе первого документа из вершины «язык» исходит только одна дуга e_5, а это слово занимает вторую позицию первого предложения $s_1=2$ (см. рис.3).

В общем случае индексный граф представляет собой очень громоздкое образование, которое требует значительных ресурсов для своего создания и хранения. Существует множество приемов, повышающих эффективность обработки таких графов. Например, для наборов документов большой мощности индексный граф удобнее строить последовательно, постепенно наращивая его размеры.

Документы набора обрабатываются в некоторой фиксированной очередности, а индексный граф пополняется новыми вершинами и дугами, отражающими предложения обрабатываемого источника. В этом случае граф претерпевает только локальные структурные изменения, он не требует масштабной перестройки, связанной с обработкой подграфов или поиском трудно вычисляемых графовых структур (например, гамильтоновых циклов, клик и др).

Мера близости документов, основанная на соответствии фраз

Представление набора документов в виде индексного графа позволяет использовать меры близости текстовых источников, основанные на сравнении целых фраз, а не отдельных слов.

Будем считать, что для некоторой пары документов d_i и d_j получен подграф вида $M_{ij} = G_i \bigcap G_j$. Пусть по этому носителю определен список общих для d_i и d_j предложений. Эффективные способы решения этой задачи обсуждались в публикациях по теории графов и анализу данных [5].

Для вычисления близости документов можно использовать следующие данные:

- число предложений, общих для документов d_i и d_j. Обозначим этот параметр через P;
- длины общих предложений $l_i, i = \overline{1, P}$;
- частоты общих предложений в обоих документах, подлежащих сравнению. Обозначим эти числа через f_{ki} и f_{kj}, $k = \overline{1, P}$, где f_{ki}, f_{rj} – частоты k-го предложения в документах d_i и d_j соответственно;
- коэффициенты важности предложений в сравниваемых документах: w_{ki}, w_{kj}.

В [17] предложена эмпирическая формула, дающая числовое значение близости двух документов d_i и d_j:

$$sim_P(d_i, d_j) = \frac{\sqrt{\sum_{k=1}^{P} [g(l_k) \times (f_{ki} w_{ki} + f_{kj} w_{kj})]^2}}{\sum_r |s_{ri}| w_{ri} + \sum_l |s_{lj}| w_{lj}} \qquad (*).$$

Здесь: $g(l_k)$ – функция, которая оценивает длину совпадающих фраз; $|s_{ri}|$, $|s_{lj}|$ – длины предложений в документах d_i и d_j соответственно. Удельный вес отдельных предложений в формуле (*) растет с ростом их длины и

частоты в обоих документах. Функция $g(l_k)$ имеет вид $g(l_k) = \left(\frac{|ms_k|}{s_k} \right)^t$, $|ms_k|$ – длина совпадающей части предложения, а t – коэффициент фрагментации $t \geq 1$. Если $t=1$, то обе части предложения (совпадающая и несовпадающая) рассматриваются и обрабатываются как независимые друг от друга. Если $t>1$, то совпадающая часть предложения вносит больший вклад в совокупное значение формулы (*).

Опыт использования мер близости, основанных на соответствии фраз, показал, что в некоторых случаях они способны оценивать сходство подобных документов как низкое. Это происходит в тех случаях, когда некоторый общий контекст можно выразить посредством одного словаря, но разной последовательностью терминов. Для кластеризации таких источников используются синтетические меры близости, которые сочетают в себе свойства мер TF×IDF и мер соответствия.

Например, синтетическая мера может быть задана следующим выражением: $sim(d_i, d_j) = \alpha \times sim_p(d_i, d_j) + (1-\alpha) \frac{\overline{d_i} \bullet \overline{d_j}}{\| \overline{d_i} \| \times \| \overline{d_j} \|}$,

где второе слагаемое представляет собой косинусную меру близости между документами, представленными в форме векторов TF×IDF, α – коэффициент смешения, принимающий значение в интервале [0,1].

Список литературы

1. Arjen van Ooyen. Theoretical aspects of pattern analysis. Режим доступа http://anc.ed.ac.uk/arjen/ (дата обращения: 20.07.2013).
2. Attributes Eui-Hong, Han George Karypis, Vipin Kumar. Min-Apriori: An Algorithm for Finding Association Rules in Data with Continuous Mining Association Rules. Режим доступа http://www-users.cs.umn.edu/~karypis/ (дата обращения: 20.07.2013).

3. Castro Félix, Vellido Alfredo, Nebot Àngela, Mugica Francisco Applying Data Mining Techniques to e-Learning Problems. Режим доступа http://sci2s.ugr.es/keel/pdf/specific/capitulo/ApplyingDataMiningTechniques.pdf/ (дата обращения: 20.07.2013)
4. Chaomei Chen. Structuring and Visualising the WWW by Generalised Similarity Analysis. Режим доступа http://www.cs.bris.ac.uk/~chen/ (дата обращения: 20.07.2013).
5. Data mining in e-learning / Romero C., Ventura S. // Witpress Boston, 2006. – 304.
6. Diday, E., Simon, J. C. Clustering analysis // In: Digital Pattern Recognition, K. S. Fu, Ed. Springer-Verlag, Secaucus, NJ. p. 47–94.
7. Douglass R. Cutting, David R. Karger, Jan O. Pedersen, John W. Tukey. Scatter/Gather: A Clustering algorithm. Режим доступа http://www.sims.berkeley.edu/~hearst/ (дата обращения: 20.07.2013).
8. Douglass R. Cutting, David R. Karger, Jan O. Pedersen. Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections. Режим доступа http://www.dsi.unive.it/~smm/docs/ (дата обращения: 20.07.2013).
9. Fern, X.Z., Brodley, C.E. Clustering ensembles for high dimensional data clustering // In Proc. International Conference on Machine Learning, 2003. p. 186-193.
10. Fred, A., Jain, A.K. Combining multiple clusterings using evidence accumulation // IEEE Tran. on Pattern Analysis and Machine Intelligence, 2005. v. 27. p. 835-850.
11. Gowda, K. C., Krishna, G. Agglomerative clustering using the concept of mutual nearest neighborhood // Pattern Recognition, 1977. v. 10. p. 105–112.
12. Jain A. K., Murty M. N., Flynn P. J. Data clustering: a review // ACM Computing Surveys, 1999. v. 31, n 3. p. 264-323.
13. Michael Steinbach, George Karypis, Vipin Kumar. A Comparison of Document Clustering Techniques. Режим доступа http://www-users.cs.umn.edu/~karypis/ (дата обращения: 20.07.2013).
14. Michalski R., Stepp R., Diday E. Automated construction of classifications: conceptual clustering versus numerical taxonomy // IEEE Trans. Pattern Anal. Mach. Intell. PAMI-5, 1983. v. 5. p. 396–409.

15. Minos N. Garofalakis, Rajeev Rastogi, Kyuseok Shim. Data Mining and the Web: Past, Present and Future. Режим доступа http://www.bell-labs.com/user/rastogi/ (дата обращения: 20.07.2013).
16. Rakesh Agrawal, Ramakrishnan Srikant. Fast Algorithms for Mining Association Rules. Режим доступа http://www.almaden.ibm.com/cs/people/ragrawal/ (дата обращения: 20.07.2013).
17. Rasmussen E. Clustering algorithms/ Режим доступа http://www.dli2.nsf.gov/ (дата обращения: 20.07.2013).
18. Salton G, Wong A, Yang C. A Vector Space Model for Automatic Indexing. // Communications of the ACM, v 18(11): p 613-620, 1975.
19. Strehl A., Ghosh J. Clustering ensembles - a knowledge reuse framework for combining multiple partitions // The Journal of Machine Learning Research, 2002. v. 3. p. 583-617.
20. Wai-chiu Wong, Ada Wai-chee. Increment Document Clustering for Web Page Classification. Режим доступа http://www.cs.cuhk.hk/~adafu/ (дата обращения: 20.07.2013).
21. Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Классификация и снижение размерности. М.: Финансы и статистика, 1989. – 450 с.
22. Дуда Р., Харт П. Распознавание образов и анализ сцен. М.: Мир, 1976. – 559 с.
23. Дюран Б., Оделл П. Кластерный анализ. М.: Статистика, 1977. – 128 с.
24. Жамбю М. Иерархический кластер-анализ и соответствия: Пер. с фр. М.: Финансы и статистика, 1988. – 342 с.
25. Журавлѐв Ю.И., Рязанов В.В., Сенько О.В. Распознавание. Математические методы. Программная система. Практические применения. М.: ФАЗИС, 2006. – 159 с.
26. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. Новосибирск: Изд. Института математики, 1999. – 270 с.
27. Классификация и кластер. / Под ред. Дж. Вэн Райзина. М.: Мир, 1980, – 390 с.