Светлов Н.М. Структуризация баз информации // Изменяющаяся Россия: аграрные преобразования начала XXI в. – результаты и перспективы. Научные труды НАЭКОР. Вып. 12. М.: Изд-во ФГНУ «Росинформагротех», 2008. — Т. 2, с.83-88

Структуризация баз информации

***Н.М. Светлов**, профессор кафедры экономической кибернетики РГАУ МСХА имени К.А. Тимирязева, д.э.н.*

В статье предложен комплекс теоретических решений проблемы структуризации семантически неоднородной информационной базы произвольной тематики. Предложенные решения нацелены на повышение эффективности поиска требуемых документов (или их реквизитов). Они сочетают высокий уровень автоматизации семантического анализа с возможностью реализации простых процедурных средств для поддержки интерактивного режима уточнения семантики документов оператором или пользователем.

Научно-технический прогресс приводит к взрывному росту объёма и всё большему тематическому разнообразию информационных ресурсов, используемых как в исследовательской деятельности, так и при подготовке управленческих решений. Отслеживать документы, содержащие требуемую информацию, становится всё труднее. Насколько доступны специалисту информационные ресурсы, создаваемые в смежных областях науки и производства, — зависит от того, успешно или нет решается задача структуризации *базы информации*, хранящей накопленные знания.

Сегодня доминирующие позиции заняла технология поиска документов, в основе которой лежит индексирование лексем. При этой технологии пользователь, желающий найти документ, относящийся к определённой предметной области, указывает ключевые слова — лексемы, которые, по его мнению, могут содержаться в искомом документе либо в определённом его реквизите. Найденные документы (если таковые имеются) ранжируются согласно некоторому эвристическому правилу, предугадывающему релевантность документа запросу пользователя. Эффективность процесса поиска во многом определяется адекватностью эвристик, заложенных в критерий ранжирования, задачам, стоящим перед пользователем базы информации.

Данный подход к поиску универсален. Его используют практически все поисковые машины глобальной вычислительной сети Интернет, многие специализированные базы документов (в частности, правовые), ряд популярных технологий поддержки принятия решений, таких, как Data mining [5]. Как правило, при достаточно развитой базе информации он даёт вполне удовлетворительные результаты. Вместе с тем точность отбора документов по ключевым терминам оставляет желать лучшего. Например, при поиске в сети Интернет документов, признаки которых не известны достоверно, доля документов, релевантных цели поиска, в общем массиве документов, выдаваемых в ответ на запрос, редко превышает 10-20%. С течением времени, в связи с увеличивающимся объёмом базы информации, этот показатель имеет тенденцию к снижению. Поиск по ключевым словам особенно неэффективен для обнаружения документов, в которых вводятся новые, заранее не известные, термины. Вышесказанное определят актуальность проблемы дальнейшего совершенствования методов структуризации информационной базы.

Альтернатива поиску по ключевым словам — использование интеллектуальных алгоритмов структуризации баз информации и поиска документов — предложена в [4]. Потребность в ней особенно остро ощущается при решении задач поиска документов нормативного, инструктивного, технического и научного содержания. Помимо повышения качества поиска и производительности труда пользователя информационной базы, её автоматизированная структуризация позволяет добиться углубления семантического анализа: единицей классификации может быть не только документ, но и его фрагмент, лишь бы он обладал информативностью, достаточной для его классификации. Последнее даёт возможность существенно повысить качество поиска требуемой информации, при котором размер «единицы обнаружения» задаётся не составителем документа, а информационной потребностью пользователя — лишь бы выполнялось естественное условие, согласно которому количество информации, содержащейся в «единице обнаружения», должно быть достаточным для её семантического анализа и классификации.

Рассмотрим методы автоматического семантического анализа, которые могут быть положены в основу создания алгоритмов структуризации баз информации. Теоретическими предпосылками автоматизированной структуризации документа являются:

- формальное описание семантики документа, основанное на теории графов [1, 2];
- представление знаний о признаках релевантности документа конкретной тематике в форме условных вероятностей [4];
- использование методов автоматического обучения.

Пусть над множеством объектов информационной базы задано отношение Q, представленное графом вида «дерево», который задаёт множество компонентов семантики анализируемых документов. Со стороны пользователя Q представляется нечётким тематическим рубрикатором, каждой вершине которого (не обязательно терминальной) соответствует множество всех документов информационной базы, упорядоченное по релевантности семантике этой вершины.

Каждой паре (q, d), где q — некоторая вершина графа Q, а d — идентификатор документа, припишем неотрицательное число p_{dq}, характеризующее схожесть данного документа с остальными документами данной рубрики, положив $\sum_{q \in Q} p_{dq} = 1 \quad \forall d$. Назовём вектор $\mathbf{p}_d = (p_{dq})$ *вектором релевантности* документа d вершинам графа Q. Он характеризует семантику документа относительно нечёткого тематического рубрикатора, представленного этим графом.

Если граф Q задан, то задача семантического анализа документа (а следовательно, и задача классификации документов) сводится к отысканию статистической оценки вектора релевантности на основе формализуемых атрибутов данного документа.

Можно предложить три взаимодополняющих подхода к оцениванию векторов релевантности.

Метод условной энтропии универсален, основан на сравнительно несложном алгоритме, но весьма требователен к вычислительным мощностям и предполагает ограничение на минимальный размер документа (обычно порядка нескольких десятков килобайт).

Распознавание образов как приём оценивания релевантности также достаточно универсально, но более требовательно к однородности структуры документов в пределах рубрики. При невыполнении этого требования возможны ошибки в оценках компонентов вектора релевантности. Подобно методу условной энтропии, распознавание образов требует высоких затрат вычислительных ресурсов.

Метод формализуемых атрибутов ограничен в своих возможностях документами специфических типов. Нечёткие правила оформления документов и их отдельных частей могут быть причиной ошибок в оценках. По сравнению с двумя предыдущими он требует значительно больших трудозатрат для проектирования и разработки, но потенциально способен обеспечить более высокую точность классификации документов и не столь ресурсоёмок.

Рассмотрим каждый из вышеназванных методов подробнее.

Метод условной энтропии основывается на использовании в качестве показателя релевантности значения подходящим образом подобранной функции вида $f(H, h, H')$, где H — количество информации [3] во всех текстах, относимых к данной рубрике с уровнем релевантности, превышающим некоторый эмпирически установленный порог (который может быть различным для разных рубрик), h — количество информации в классифицируемом тексте, H' — количество информации во всех текстах данной рубрики после её пополнения классифицируемым документом. В простейшем случае функция оценивания коэффициента релевантности может иметь вид $1 - (H' - H) / h$. Её значение изменяется в диапазоне $0 < f(H, h, H') \le 1$. Чем ближе её значение к 1, тем данный документ более схож с документами, уже присутствующими в рубрике.

Предметным аналогом понятия количества информации по Колмогорову, которое здесь используется, является не число символов в оригинальном документе, а его размер после оптимального кодирования с использованием наилучшего доступного алгоритма.

Величина H может быть определена с учётом релевантности входящих в неё документов. Например, при наличии достаточно репрезентативной рубрики можно оценивать количество содержащейся в ней информации методом случайной выборки семантически цельных фрагментов (например, абзацев) из относящихся к ней документов, причём доля выборки в объёме конкретного документа (измеренная по числу символов оригинального текста) пропорциональна его релевантности.

При использовании метода распознавания образов на входные синапсы нейронной сети, обученной отличать от других документы, относящиеся к «своей» рубрике, подаются сигналы, формализующие лексические единицы документа. Результатом работы нейронной сети является сигнал, интенсивность которого находится в достаточно тесной положительной корреляции с мерой релевантности документа данной рубрике. В качестве обучающей выборки для настройки нейронной сети используется текущее содержимое рубрики. Требование к числу входящих в неё документов в этом случае более жёсткое, чем в предыдущем.

Метод формализуемых атрибутов основывается на представлении знаний о семантике документов в форме набора продукционных правил, отображающих атрибуты документов на логическое значение (возможно, нечёткое). Множество логических значений, полученных при посредстве всей совокупности правил, отображается на меру релевантности. Если последняя допускает интерпретацию в терминах теории вероятности и заданы априорные вероятности соответствия документа рубрике, то искомый вектор релевантности может быть определён применением формулы Байеса к заданным априорным вероятностям. Меры релевантности различным рубрикам, полученные с помощью продукционных правил, подставляются в формулу Байеса в качестве условных вероятностей.

Как только оценки векторов релевантности для всех документов информационной базы получены, процедура поиска требуемых документов (или реквизитов) становится намного проще и эффективнее. Для получения требуемых документов пользователь информационной базы может выбрать, по крайней мере, один из двух приёмов:

♦ для отбора требуемых документов — сформулировать условия, которым должен отвечать вектор релевантности искомого документа, в форме неравенств;

♦ для ранжирования документов информационной базы по критерию, отражающему цели поиска, — задать вектор весовых коэффициентов, применяемых к компонентам вектора релевантности. В последнем случае документы ранжируются по показателю $\mathbf{wp}_d$, где **w** — определённый пользователем вектор весов. При решении простейших задач поиска этот вектор может содержать единственный ненулевой компонент.

Данный метод поиска имеет и серьёзный недостаток: пользователь не может сформулировать запрос, семантика которого не может быть представлена выпуклой линейной комбинацией семантических компонентов, соответствующих вершинам графа *Q*. Другими словами, обращаясь к подобной системе, пользователь ограничен в возможностях уточнения своего запроса множеством рубрик, которые имеются в данном тематическом рубрикаторе. Однако этот недостаток преодолевается совмещением поиска по векторам релевантности с традиционным поиском по ключевым словам, а также постепенным совершенствованием и развитием самого рубрикатора.

Рассмотренные выше методы оставляют нерешёнными две практические проблемы: наполнение новой рубрики, которая ещё не содержит документов-образцов, и опасность «эволюционирования» семантики рубрики под влиянием систематических особенностей вновь включаемых в неё документов.

Первая проблема требует привлечения живых экспертов-операторов, которые должны определить начальную совокупность документов, релевантных новой рубрике. Заметим, что если рубрику можно определить как выпуклую линейную комбинацию семантики существующих рубрик, в её создании нет смысла: ведь в этом случае достаточно соответствующим образом сформулировать запрос к информационной базе. Поэтому автоматизированные методы подбора документов в новые рубрики едва ли могут быть предложены.

Вторая проблема также требует вмешательства живых экспертов. Если задача создания новой рубрики решается сравнительно редко и не требует излишне больших трудозатрат, то выявление ошибочно классифицированных документов чрезвычайно трудоёмко. Но в данном случае можно воспользоваться помощью пользователей информационной базы, которых целесообразно наделить правом помечать документы (реквизиты), релевантность которых данной рубрике оценена, на их взгляд, неверно. Окончательное заключение о релевантности документов (реквизитов), на которые указали пользователи, остаётся, как правило, за администратором информационной базы, который может привлечь для решения

выявленной проблемы эксперта, заслуживающего доверия. Но в частных случаях, если круг пользователей информационной базы ограничен (например, так обстоит дело с информационными базами научного или внутрифирменного пользования), все пользователи или их часть могут быть наделены правом менять уровень релевантности любого документа любой рубрике либо только тем рубрикам, которые соответствуют квалификации пользователя.

Введение новой рубрики должно сопровождаться перевычислением векторов релевантности всех документов базы. Целесообразно также периодически перевычислять все векторы релевантности, чтобы учесть последствия пополнения информационной базы новыми документами и корректировки, внесённые вручную.

Библиографический список

1. Алфёрова З.В. Математическое обеспечение экономических расчётов с использованием теории графов. М.: Статистика, 1974. — 208 с

2. Землянский А.А., Светлов Н.М. Теоретические основы формализации линейного экономико-математического моделирования // Современные информационные технологии в экономике: Сборник научных трудов / Моск. эконом.-стат. ин-т. М., 1992. — С. 85-100.

3. Колмогоров А.Н. Три подхода к определению понятия «количество информации» // Проблемы передачи информации. – 1965. – Т.1, №1. – С.3-11.

4. Светлов Н.М., Светлова Г.Н. Применение искусственного интеллекта в информационных технологиях. М.: Изд-во МСХА, 2004.

5. Berry M., Linoff G. Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management: 2nd edition. Wiley Computer Publishing, 2004. — 672 p.