УДК 004.9

А. М. Федотов [1], **В. Б. Барахнин** [2]

Институт вычислительных технологий СО РАН
пр. Акад. Лаврентьева, 6, Новосибирск, 630090, Россия

Новосибирский государственный университет
ул. Пирогова, 2, Новосибирск, 630090, Россия
E-mail: [1] fedotov@nsu.ru, [2] bar@ict.nsc.ru

К ВОПРОСУ О ПОИСКЕ ДОКУМЕНТОВ «ПО АНАЛОГИИ» *

Статья посвящена анализу формализации понятий аналогии и сходства, а также изложению процедуры поиска документов «по аналогии».

Ключевые слова: информационный поиск, аналогия, сходство, кластеризация.

Введение

В свое время Альберт Эйнштейн предупреждал, что до конца XX в. человечество станет свидетелем «информационного взрыва». Сегодня можно утверждать, что этот «взрыв» произошел, причем сила и скорость его «ударной волны» растет с каждым днем. Причина его возникновения связана не столько с гигантским прогрессом в информационных технологиях, сколько с возросшим во много раз потоком информации, необходимой для жизни современного общества. Тем не менее, как неоднократно повторял Козьма Прутков, «нельзя объять необъятное», поэтому один из наиболее распространенных подходов к удовлетворению информационных потребностей заключается в поиске документов, которые в том или ином смысле аналогичны документу (или множеству документов), уже известному данному лицу.

Напомним, что аналогия означает сходство предметов (явлений, процессов и т. д.) в каких-либо свойствах. При умозаключении по аналогии знание, полученное из рассмотрения какого-либо объекта («модели»), переносится на другой, менее изученный (менее доступный для исследования, менее наглядный и т. п.) в каком-либо смысле, объект. По отношению к конкретным объектам заключения, получаемые по аналогии, носят, вообще говоря, лишь вероятный характер [1].

Методика умозаключений по аналогии, изложенная, например, в [2], состоит в следующем. Если более изученному объекту A присущи свойства P_1, P_2, …, P_n, P_{n+1}, а изучение менее изученного объекта B показало, что ему присущи свойства P_1, P_2, …, P_n, то можно сделать предположение о наличии у объекта B свойства P_{n+1}. Степень вероятности правильного умозаключения по аналогии будет тем выше, чем: 1) больше известно общих свойств у сравниваемых объектов; 2) насколько существенны обнаруженные у них общие свойства; 3) насколько глубже познана взаимная закономерная связь этих сходных свойств. Для повышения надежности умозаключения по аналогии желательно, чтобы общие свойства P_1, P_2,..., P_n были возможно более специфичными для сравниваемых объектов, а свойство P_{n+1}, напротив, должно быть наименее специфичным.

Важно подчеркнуть, что в приведенном выше определении понятия *аналогия* речь идет лишь о *сходстве* сравниваемых объектов, поскольку устанавливать *тождество* возможно

* Работа выполнена при частичной поддержке РФФИ (проекты № 07-07-00271, 08-07-00229, 09-07-00277), президентской программы «Ведущие научные школы РФ» (грант № НШ-931.2008.9) и интеграционных проектов СО РАН.

лишь при высокой степени абстрактности определения рассматриваемых свойств, что весьма затруднительно в конкретных исследованиях. Однако сразу же возникает вопрос: а что же именно следует понимать под *сходством* некоторых объектов (в частности, документов)?

Общелитературное толкование слова *сходство*, приведенное, например, в словаре Ушакова: «Одинаковость, подобие, соответствие в чем-нибудь с кем-чем-нибудь» [3], с формальной точки зрения носит характер *circulus vitiosus*, поскольку в том же словаре *одинаковость* определяется как «…тождество, совпадение чего-нибудь», *подобие* – как «образ чего-нибудь, нечто похожее, сходное», *соответствие* – как «соотношение между чем-нибудь, выражающее согласованность, равенство», *равенство* – «одинаковость, полное сходство» и т. д. Сказанное, разумеется, не является упреком в адрес составителей словаря, которые не ставили перед собой цель построения «формально-аксиоматической» модели семантики русского языка, но служит свидетельством того, что *сходство* относится к числу «базисных» слов, понимание которых априори предполагается для любого носителя языка.

Таким образом, рассматривая далее основные постановки задач информационного поиска по аналогии, мы вправе подразумевать общелитературное понимание слова *сходство,* памятуя, однако, о том, что в основе алгоритмов автоматизации нахождения «похожих» документов должно лежать более формализованное толкование термина, которое будет обсуждаться в следующем разделе статьи.

Разумеется, поиск документов по аналогии интересует нас, прежде всего, в плане возможности его автоматизации. Наиболее простыми для автоматизированного исследования (т. е. для установления посылок умозаключений по аналогии) свойствами документов являются их выходные (библиографические) данные, а также совокупность слов и их последовательностей, входящих в аннотацию и полный текст документов. Напротив, автоматизация непосредственного определения содержания (и даже тематики) документа, выходные данные которого не содержат классификационных признаков, весьма затруднительна, поэтому именно эти свойства целесообразно устанавливать по аналогии. Важно опять-таки подчеркнуть, что речь идет, естественно, не о *тождестве* содержания документов, а лишь о его *сходстве*.

Рассмотрим несколько практических проблем, решаемых путем поиска документов по аналогии.

Поиск научных публикаций, схожих с уже известными. Ежегодно в мире публикуются миллионы научных статей, постольку даже в узкоспециализированных отраслях науки просматривать весь объем информации практически невозможно. Вследствие этого широкое распространение получили электронные носители информации о новых научных публикациях, например:

- базы данных реферативных журналов;
- базы данных Current Contents;
- специализированные сетевые базы данных типа Zentralblatt MATH.

Однако следует отметить, что информационные потребности научных работников, когда они в процессе исследования находятся на этапах изучения уже имеющихся в данной области результатов и научного поиска, характеризуются невысокой четкостью осознания и выражения (см., например, [4]). Имеет место ситуация, отраженная в сюжете русской народной сказки «Пойди туда – не знаю куда, принеси то – не знаю что», причем акцент ставится на второй части фразы, поскольку известно, что описания документов, относящихся к той или иной научной тематике, заносятся в упомянутые реферативные базы данных.

С другой стороны, у каждого исследователя за годы его работы образуется «картотека» библиографических описаний статей, книг и т. д., представляющих для него интерес. Основной критерий их отбора – личные интересы исследователя. В настоящее время такие картотеки хранятся, как правило, на электронных носителях, что позволяет организовывать интегрированные картотеки путем объединения ресурсов совместно работающих исследователей. Так как документы в такие картотеки заносятся из различных источников, имеющих, вообще говоря, различные классификаторы, а иногда и не имеющих их вообще (что характерно, например, для многих научных журналов), то поиск и классификация документов по формальным классификационным признакам зачастую невозможны.

Таким образом, один из основных подходов к процессу отбора публикаций, которые могут представлять интерес для конкретного исследователя или группы совместно работающих

исследователей, заключается в нахождении по данному документу (или множеству документов) класса документов, аналогичных по содержанию. В качестве информационного запроса предполагается задание непустого множества документов, а в качестве результата выполнения запроса выдаются документы, каждый из которых в определенном смысле близок к одному из документов, входящих в заданное множество. При этом надо иметь в виду, что при интеграции нескольких баз данных можно столкнуться с наличием в объединенном множестве дубликатов одного и того же документа, которые для удобства пользователя следует исключать из окончательных результатов поиска. Тем самым мы сталкиваемся с ситуацией, когда следует вводить ограничения на «излишнее» сходство (в данном случае тождество) находимых документов.

Отметим, что в ставшей уже классической монографии Дж. Солтона [5] для сравнения статей из научных журнала с целью определения их сходства предлагается использовать атрибуты их библиографических описаний, такие как: авторы, название журнала, а также количественные признаки – частота терминов в аннотации и заглавии статьи.

Объединение новостных сообщений в сюжеты. В настоящее время в Рунете существуют несколько десятков новостных информационных порталов, основной принцип работы которых заключается в аккумулировании новостной информации, публикуемой на сайтах информационных агентств, и объединении сообщений, освещающий ход развития того или иного события, в так называемые *сюжеты* (такое объединение обычно реализуется посредством публикации в конце сообщения гиперссылок на другие сообщения, относящиеся к этому же сюжету). Огромные объемы поступающей информации требуют автоматизации процесса выявления сообщений сходной тематики, причем, как и в задаче поиска научных публикаций, здесь также возникает проблема удаления избыточной информации. Однако в данном случае дело осложняется тем, что в отличие от предыдущей задачи, при формировании новостных сюжетов требуется удалять не только полные, но и нечеткие дубликаты, возникающие, например, вследствие того, что разные информационные агентства независимо друг от друга сообщили одну и ту же новость (естественно, несколько разными словами).

Поиск художественных произведений. Одним из основных критериев, которым руководствуется большинство читателей при выборе произведений художественной литературы, является сходство выбираемых книг с книгами, ранее понравившимися данному читателю. Так как количество наименований художественных книг постоянно растет, то сделать выбор «классическим» путем, просматривая книги в магазине или библиотеке, становится все более затруднительно. И здесь на помощь также могут прийти современные информационные технологии, поскольку Интернет содержит большое количество документов, непосредственно представляющих произведения художественной литературы (в виде их полных текстов) или достаточно подробно описывающих такие произведения (в виде выходных данных и аннотаций книг, продаваемых в интернет-магазинах).

Особенности алгоритмов для каждой из этих задач мы рассмотрим в конце статьи, после того, когда будет описана формализация понятия *сходства* и приведены подходы к количественной оценке степени сходства объектов.

Формализация понятия *сходства*

Начнем с определения сходства, данного в Философском энциклопедическом словаре (ФЭС) [6]: «Сходство, отношение, родственное равенству. При наличии у пары объектов хотя бы одного общего признака можно говорить о сходстве объектов этой пары. Ввиду многообразия признаков на одной паре могут индуцироваться разные отношения сходства, а ввиду повторяемости признаков – одно отношение сходства на разных парах. Отношения сходство на разных парах интенсионально [1] совпадают, если они определены одним и тем же признаком. Если этот признак – четко определенное свойство, присущее каждому элементу рассматриваемых пар, то отношение сходства всегда рефлексивно, симметрично и транзитивно,

[1] Интенсиональность – зависимость истинности высказываний не только от истинности составляющих их более простых высказываний, но и от психологических, прагматических и модальных оттенков смысла этих высказываний.

т. е. совпадает с отношением равенства (эквивалентности) на множестве объектов, входящих во все такие пары. Если же этот признак – отношение, то сходство по отношению, вообще говоря, рефлексивно, но не обязательно транзитивно или даже симметрично…»

Какие же важные характеристики понятия сходства содержатся в этом определении?

Во-первых, отмечается интенсиональность отношения сходства, т. е. его контекстная зависимость, что фактически в той или иной мере обуславливает субъективность или, по крайней мере, интерсубъективность рассматриваемого отношения. Особо подчеркнем, что в период публикации ФЭС в отечественной гносеологии безраздельно господствовала ленинская теория отражения [7], согласно которой процесс восприятия объектов внешнего мира и их свойств носит объективный характер: «…логично предположить, что вся материя обладает свойством, по существу родственным с ощущением, свойством отражения» [Там же. С. 91]. Тем не менее, применительно к восприятию отношения сходства даже тогда нельзя было не отметить элемент субъективности (который, впрочем, в той или иной степени присущ всему процессу познания: «Информация всегда относительна, она зависит… от того, какой информационной системой она воспринимается» [8]). Житейской иллюстрацией интенсиональности сходства может служить не вышедший на экраны эпизод из сценария известного советского фильма «Мимино» (1977, режиссер Г. Данелия), когда главные герои фильма Валико (исполнитель роли В. Кикабидзе) и Рубик (исполнитель роли Ф. Мкртчян) входят в лифт гостиницы, где стоят два японца, похожие друг на друга, с точки зрения советского (притом отнюдь не только русского!) зрителя, как близнецы. Увидев входящих, японцы говорят друг другу по-японски (с русскими субтитрами): «Как все эти русские похожи друг на друга!».

Во-вторых, в ФЭС подчеркивается, что отношение сходства, определяемое *всего лишь одним признаком*, далеко не всегда обладает свойствами транзитивности и даже симметричности (проводимое в известной книге «Равенство, сходство, порядок» [9] подробное обсуждение нетранзитивности отношения сходства не столь показательно, поскольку основано на использовании *разных* признаков при сравнении разных пар объектов). Пример нетранзитивного сходства довольно очевиден: вполне естественно (в соответствующих контекстах) звучит как выражение «четыре часа ночи» (ср. «час ночи»), так и выражение «четыре часа утра» (ср. «семь часов утра»), т. е. временной признак события «произойти в 4.00» сходен (в зависимости от контекста) в смысле принадлежности к одному и тому же времени суток как с признаком «произойти в 1.00», так и с признаком «произойти в 7.00». Вместе с тем невозможно назвать время суток, к которому одновременно можно было бы отнести события, происшедшие в 1.00 и в 7.00. Случаи несимметричного сходства не столь тривиальны, и их обсуждению мы посвятим отдельный пункт.

Из определения ФЭС, однако, неясно, как конкретно можно установить сходство между объектами. Некоторую ясность вносит определение сходства из Большой советской энциклопедии (БСЭ) [10]: «…соответствие отображения, образа своему оригиналу… Оно включает три основные отношения: соответствие качественных характеристик отображения особенностям оригинала (например, ощущение зеленого цвета листьев растения соответствует определенной длине электромагнитных волн, излучаемых поверхностью листьев); соответствие структур отображения структурам оригинала (например, структура географической карты соответствует геометрическим структурам местности), причем разные виды соответствия структур могут описываться с помощью различных математических отображений – изоморфизма, гомоморфизма и др.; соответствие количественных характеристик отображения и оригинала (например, количественные значения состояний термостата соответствуют измеряемой температуре тела)».

Важный момент, отмеченный в определении БСЭ, – представление процедуры установления сходства как сравнения объекта с неким «оригиналом». Очевидно, что такой «оригинал» далеко не всегда имеет отношение к реальному происхождению сравниваемых объектов (так, внешнее сходство двух людей, заведомо не являющихся близкими родственниками, обусловлено отнюдь не наличием в обозримом прошлом их гипотетического общего предка). Тем самым в качестве «оригиналов (прежде всего, при установлении сходства по качественным характеристикам) выступают так называемые общие понятия (говоря философским языком, универсалии). Мы не будем здесь обсуждать вопрос о природе универсалий, являющийся, начиная с античности, предметом острейших философских споров (см., например, [11] и

имеющуюся в статье библиографию). Отметим лишь, что наличие универсалий, рассматриваемых, по крайней мере, как некие общепринятые обозначения, позволяет рассматривать множество универсалий, с помощью которых описывается то или иное свойство объекта, в качестве *измерительной шкалы*. Известно (см., например, [12]), что шкалы, по которым устанавливается сходство, бывают *номинальные* (задающие только наименования значений отслеживаемых свойств), *порядковые* (на множестве значений свойств задано отношение порядка, но отсутствует возможность количественного сравнения значений) и *арифметические* (множество значений которых допускает ту или иную форму количественного сравнения). При этом набор возможных значений неарифметических шкал необязательно полностью задан априори: например, при изучении документов множество значений шкалы «авторы» может пополняться, если автор нового документа впервые фигурирует в таком качестве для данной коллекции документов. Кстати, на примере этой же шкалы можно видеть, что некоторые свойства того или иного объекта могут иметь сразу несколько различных значений (в данном случае – когда у документа несколько авторов).

Определение БСЭ интересно еще и тем, что в нем четко выделяются два типа сходства: по качественным и по количественным характеристикам (правда, пример, иллюстрирующий соответствие качественных характеристик, подобран, на наш взгляд, не совсем удачно: в нем соответствие устанавливается фактически по количественной характеристике – длине волны). Что же касается сходства по структурным характеристикам, то оно сводится к качественному (а иногда, при необходимости уточнения совпадающих качественных характеристик, и к количественному) соответствию структурных элементов сравниваемых объектов. Но именно структурное сходство наиболее важно для возможности делать умозаключения по аналогии. Вот что об этом писал известный американский математик венгерского происхождения Д. Пойа в книге «Математика и правдоподобные рассуждения» [13]: «Рассматривая в музее естественной истории скелеты различных млекопитающих, вы можете обнаружить, что все они страшны. Если в этом все сходство, которое вы между ними обнаружили, то вы видите не такую уж сильную аналогию. Однако вы можете подметить удивительно много говорящую аналогию, если рассмотрите руку человека, лапу кошки, переднюю ногу лошади, плавник кита и крыло летучей мыши – эти столь различно используемые органы, как состоящие из сходных частей, имеющих сходное отношение друг к другу. Последний пример иллюстрирует наиболее типичный случай выясненной аналогии; *две системы аналогичны, если они согласуются в ясно определенных отношениях соответствующих частей*».

Таким образом, начальный этап формализации процедуры поиска объектов «по аналогии» состоит в выделении основных составных частей или свойств, присущих данному типу объектов, и задании соответствующих шкал путем установления множества возможных значений каждого из этих свойств.

Проведенный анализ значений этого термина (и родственных ему понятий) может, на первый взгляд, показаться излишне подробным. Однако без такого анализа практически невозможно обоснование адекватности выбора той или иной формальной модели сходства, так как при поиске документов «по аналогии» оценка релевантности документа носит интенсиональный, а зачастую и весьма субъективный характер (напомним, что *релевантными* называются документы, содержание которых соответствует информационному запросу), поскольку такой поиск допускает произвол в выборе элементов структуры, по которым устанавливается сходство, а также, как мы увидим в дальнейшем, в способе задания количественной оценке степени (меры) сходства, в установлении ее порогового значения, отделяющего «похожие» документы от «непохожих» и т. п.

О несимметричном сходстве

В приведенном выше определении сходства из БСЭ отмечено, что если признак, по которому определяется сходство, является отношением, то такое сходство может быть несимметричным. В качестве иллюстрации этого утверждения приведем следующий пример. На множестве городов – административных центров субъектов Российской Федерации установим в качестве признака сходства близость расстояний между городами. С помощью географической карты нетрудно увидеть, что для Калининграда наиболее близкими (а следовательно, и

сходными в заданном понимании) городами будут Псков и Смоленск. Однако для Пскова в качестве «наиболее сходных» городов будут определены отнюдь не Калининград, а Новгород и Санкт-Петербург, для Смоленска – Брянск и Калуга.

Приведенный пример несимметричного сходства построен на том, что отношение сходства определялось на достаточно широком множестве объектов. Однако гораздо более интересно рассмотреть вопрос о том, может ли сложиться такая ситуация, когда при наличии пары объектов (*A* и *B*) мы сделаем вывод о том, что *A* похож на *B*, но при этом *B* не похож на *A*?

На первый взгляд, возможность того, что сходство двух *отдельно взятых* объектов бывает несимметричным, кажется оксюмороном или, по меньшей мере, парадоксом. Именно так воспринималось известное еще, по-видимому, с античных времен высказывание «Помпей и Цезарь очень похожи, особенно Цезарь» (для дальнейших рассуждений важно отметить, что современники событий отнюдь не признавали кажущееся нам почти бесспорным превосходство Цезаря. Как писал Плутарх («Сравнительные жизнеописания. Помпей», гл. I), «никто из римлян, кроме Помпея, не пользовался такой любовью народа, – любовью, которая возникла бы столь рано, столь стремительно возрастала в счастье и оказалась бы столь надежною в несчастьях»).

Особую остроту тема несимметричного сходства обрела в христианском богословии в связи с библейскими стихами «сотворил Бог человека по образу своему» (Быт. 1, 27) и «когда Бог сотворил человека, по подобию Божию создал его» (Быт. 5, 1), а также с многочисленными высказываниями Отцов Церкви о том, что Антихрист, явившись в мир, будет стремиться во всем походить на Христа. Мысль о том, что имеет место и «обратное» сходство (которая следует из обыденного понимания сходства) казалось многим в первом случае, как минимум, вольнодумством, а во втором – откровенным кощунством. Однозначного разрешения эта коллизия не получила даже в отдельной взятой католической традиции. Так, французский философ и публицист граф Жозеф де Местр (Joseph Marie de Maistre), долгое время проживший в России в качестве посланника сардинского короля, в книге «Санкт-Петербургские вечера» (изд. 1821) так обосновывал возможность несимметричного сходства: «Сходство между человеком и его Создателем есть сходство изображения и образца… Если же кто-то считает, будто мы говорим, что *человек похож на свой портрет*, то нелепостью этой он обязан самому себе, ибо мы утверждаем прямо противоположное».

Тем не менее уже в XX в. находились мыслители, продолжавшие отстаивать (применительно к рассматриваемой коллизии) приоритет логической связи сходства и тождества. Например, известный английский писатель и христианский мыслитель Г. К. Честертон (Gilbert Keith Chesterton) в эссе «Франциск Ассизский» заметил: «В одной из своих блестящих полемических работ кардинал Ньюмен (John Henry Newman) обронил фразу, которая может служить примером смелости и логической ясности католичества. Рассуждая о том, как легко принять истину за нечто противное ей, он говорит: “Если Антихрист похож на Христа, то и Христос, наверное, похож на Антихриста”. Религиозному чувству неприятен конец этой фразы, но опровергнуть ее может лишь тот, кто сказал, что Помпеи и Цезарь очень похожи, особенно Цезарь».

Почему же так трудно признать возможность несимметричного сходства двух объектов? Ответ на этот вопрос, по-видимому, вытекает из приведенных рассуждений Г. К. Честертона: отношение сходства нередко воспринимается как некое обобщение (а то и полный аналог, как в процитированных выше определениях из словаря Ушакова) отношения тождества, которое, безусловно, симметрично с точки зрения классической логики. При этом, однако, не учитывается, что общее понятие может и не обладать некими свойствами частного.

Интересно отметить, что в русской художественной литературе XX в. возможность несимметричного сходства признавалась вполне допустимой (хотя и воспринималась как любопытный феномен). Так, известный русский писатель В. А. Солоухин в эссе «Третья охота» отмечал: «…я, много раз принимавший издали валуи за белые грибы, хочу сказать, что ни разу еще, увидев настоящий белый гриб, я не принял его за валуй. У Глазкова [2] есть четверо-

[2] Николай Иванович Глазков – русский поэт, основоположник «самиздата» (1939 г.).

стишие о необратимости сравнения. Там говорится о том, что свистящий на плите чайник напоминает сирену, но настоящая сирена не напоминает свистящий чайник. Так и здесь».

Приведенные примеры несимметричного сходства позволяют провести формальное описание тех ситуаций, в каких может наблюдаться этот вид сходства. Именно свойства объектов (зачастую носящие сложный, «комплексный» характер) устанавливаются с помощью порядковой шкалы. При этом признается, что объект, обладающий свойством с меньшим значением, сходен с объектом, обладающим свойством с бóльшим значением, но обратного сходства может, вообще говоря, и не быть.

Применительно к поиску документов указанная ситуация может наблюдаться, например, при установлении сходства между кратким изложением документа (будь то реферат научной статьи, детское издание «Приключений Гулливера» Дж. Свифта или «краткий пересказ» «Войны и мира» Л. Н. Толстого) и его полной версией. В большинстве случаев, имея краткое изложение заинтересовавшего его документа, пользователь считает целесообразным найти его полную версию. Напротив, пользователь, имеющий полный текст научной публикации, вряд ли будет рассматривать в качестве пертинентного (т. е. соответствующего информационной потребности) результата поиска сходных с ней документов реферат этой же публикации. Аналогично читатель, которому понравилось классическое произведение художественной литературы, вряд ли захочет найти его «детское издание (если, конечно, такая цель не ставится специально; но применительно и к этой ситуации следует отметить, что знаменитый английский писатель, профессор филологии Оксфорда Дж. Р. Р. Толкин (John Ronald Reuel Tolkien) в эссе «О волшебных сказках» достаточно негативно отзывался о тенденции создания «смягченных» обработок классических сказок). Наконец, использование «кратких пересказов» классики является, на наш взгляд, делом малоэтичным, и, следовательно, ответственный разработчик информационно-поисковой системы вправе поставить ограничение на возможность удовлетворения такого рода «информационных потребностей».

Определение меры близости между объектами

Вернемся к изложению процедуры поиска объектов «по аналогии». Итак, пусть мы уже выделили основные свойства, присущие данному типу объектов, и задали подходящие шкалы, описывающие множества возможных значений каждого из свойств (если рассматриваемые объекты – документы, то в качестве шкал для определения меры сходства обычно используются атрибуты библиографического описания документов). Далее необходимо провести нормализацию шкал, введя на каждой из них «частную» меру сходства (иногда называемую нормативной операцией сопоставления двух значений свойства Ψ_i), т. е. функцию, заданную на множестве значений i-й шкалы Ψ_i (а фактически – на множестве сравниваемых объектов D) следующим образом:

$$m_i : \Psi_i \times \Psi_i \rightarrow [0, 1],$$

причем функция m_i в случае полного сходства принимает значение 1, в случае полного различия – значение 0.

Процедура нормализации зависит от типа шкалы (подробнее см. [12]). Так, для арифметических и порядковых свойств Ψ_i на множестве их значений $\{\psi_i^n\}$ (здесь $\psi_i^n = \psi_i(d_n)$, где n – номер объекта) всегда существуют минимальное ψ_i^* и максимальное ψ_i^{**} значения. Тогда для арифметических свойств можно положить

$$m_i(\psi_i^k, \psi_i^l) = \frac{|\psi_i^k - \psi_i^l|}{\psi_i^{**} - \psi_i^*},$$

а для порядковых

$$m_i(\psi_i^k, \psi_i^l) = \frac{\Delta n_i^{k,l} + 1}{\Delta n_i^{*,**} + 1},$$

где $\Delta n_i^{k,l}$ – число различных значений ψ_i^n, лежащих между ψ_i^k и ψ_i^l (если $\psi_i^k = \psi_i^l$, то $\Delta n_i^{k,l}$ полагается равным –1), а $\Delta n_i^{*,**}$ – число различных значений ψ_i^n, лежащих между ψ_i^* и ψ_i^{**}.

Наконец, для номинальных шкал мера сходства определяется следующим образом: если значения свойств объектов совпадают, то мера близости по этой шкале равна 1, иначе 0. При этом необходимо учитывать, что значения свойств объектов для номинальной шкалы могут быть составными (например, документ может иметь сразу нескольких авторов). В таком случае $m_i = n_{i1} / n_{i0}$, где $n_{i0} = \max\{n_{i0}(d_1), n_{i0}(d_2)\}$, а $n_{i0}(d_j)$ – общее количество элементов, составляющих значение i-го свойства объекта d_j, n_{i1} – количество совпадающих элементов.

После того, как подсчитаны мера сходства по каждой из шкал, можно приступить к вычислению меры сходства $m(d_1, d_2)$ между объектами, входящими в заданное множество, и объектами, среди которых мы ищем аналогичные заданным. Для этого обычно используется одна из стандартных формул вычисления расстояний с весовыми коэффициентами, чтобы вычисленное значение меры не превосходило 1. Весовые коэффициенты (они, разумеется, неотрицательны) в простейшем случае равны между собой, однако путем задания весовых коэффициентов, отличных друг от друга, мы можем указать априорную относительную важность шкал. Более того, значения весовых коэффициентов могут определяться и предполагаемой апостериорной достоверностью данных соответствующей шкалы, т. е. в определенных случаях один из коэффициентов может быть увеличен с пропорциональным уменьшением остальных. Например, полное (или даже «почти полное») совпадение значений атрибута «авторы» документа d_1 и документа d_2 более весомо в случае, когда количество значений этого атрибута в документе d_1 достаточно велико (по сравнению со случаем, когда документ d_1 имеет всего одного автора).

Использование для вычисления меры сходства между объектами d_1 и d_2 стандартной евклидовой метрики

$$m_E(d_1, d_2) = \sqrt{\sum (a_i m_i(d_1, d_2))^2}, \quad \text{где } \sum a_i^2 = 1, \tag{1}$$

оказывается не всегда удобным из-за заметного влияния отдельных больших значений m_i. Этот недостаток менее заметен при использовании, например, расстояния Хемминга

$$m_H(d_1, d_2) = \sum a_i \, m_i(d_1, d_2), \quad \text{где } \sum a_i = 1. \tag{2}$$

Напротив, если понимание сходства для конкретной задачи подразумевает отсутствие больших различий по любой отдельно взятой шкале, то целесообразно использовать расстояние Чебышева

$$m_\infty(d_1, d_2) = \max |a_i \, m_i(d_1, d_2)|, \quad \text{где } \max a_i \le 1. \tag{3}$$

В формулах (1)–(3) выражение $m_i(d_1, d_2)$ означает $m_i(\psi_i(d_1), \psi_i(d_2))$.

Установление аналогии

Для непосредственной процедуры нахождения объектов, аналогичных объектам из заданного множества, необходимо задать пороговое значение меры сходства $r \in (0, 1)$. Если заданное множество D_* состоит из одного объекта d_*, то при $m(d_*, d_j) \le r$ делается вывод, что объект d_j аналогичен заданному, в противном случае считается, что аналогия отсутствует. Ситуация осложняется, если множество D_* содержит более одного объекта. Тогда критерием аналогичности объекта d_j элементам множества D_* служит неравенство $m(D_*, d_j) \le r$, в котором $m(D_*, d_j)$ – расстояние от объекта d_j до множества D_* (обычно под этим подразумевается

минимум расстояний от объекта d_j до элементов множества D_*, хотя иногда в качестве $m(D_*, d_j)$ целесообразно рассматривать расстояние от объекта d_j до определенного тем или иным способом «центра» множества D_*). Независимо от количества элементов в множестве D_* возможно задание «градаций аналогичности», определяемых посредством набора чисел $\{r_i\}$, $i = 1, ..., n$, где $r_k < r_l$ при $k < l$. Если $r_k < m(D_*, d_1) \le r_{k+1}$, а $r_l < m(D_*, d_2) \le r_{l+1}$ при $k < l$, то считается, что объект d_1 более схож с элементами множества D_*, чем объект d_2. Введение градаций аналогичности используется, например, для установления приоритета просмотра документов, найденных в процессе информационного поиска.

Указанные процедуры поиска аналогичных документов могут быть снабжены дополнительными условиями, связанными, например, с исключением из поисковой выдачи соответствующих документов при реализации описанной выше ситуации несимметричного сходства.

Несколько иной подход к нахождению аналогичных объектов связан с кластеризацией объектов объединенного множества, включающего в себя как элементы множества D_*, так и объекты, относительно которых необходимо установить наличие или отсутствие аналогии с элементами множества D_* (напомним, что кластеризацией называется разбиение множества объектов на классы, при котором элементы, объединяемые в один класс, имеют большее (в определенном смысле) сходство, нежели элементы, принадлежащие разным классам). При этом объектами, аналогичными элементам множества D_*, признаются объекты, принадлежащие классам, содержащим определенное количество элементов D_* (это количество может быть задано как абсолютная величина или как доля элементов D_* в данном классе).

Подробный обзор алгоритмов кластеризации содержится, например, в монографии [14]. Сравнение нескольких алгоритмов кластеризации применительно к задаче установления сходства документов сделано в работе [15].

Об оценке эффективности поиска «по аналогии»

Как уже отмечалось, при поиске документов «по аналогии» оценка релевантности, а тем более пертинентности документа носит интенсиональный, а зачастую и весьма субъективный характер (напомним, что *релевантными* называются документы, содержание которых соответствует информационному запросу), поскольку процедура поиска допускает произвол в выборе элементов структуры, по которым устанавливается сходство, в способе задания количественной меры сходства, в установлении ее порогового значения, отделяющего «похожие» документы от «непохожих» и т. п. К тому же остается практически неустранимая зависимость результата поиска «по аналогии» от всей совокупности документов, входящих в информационный массив, по которому осуществляется поиск. Попросту говоря, вывод о схожести объекта «кошка» с объектом «корова» различается в случае, когда «информационный массив» есть множество «лев, корова», и в случае, когда «информационный массив» – корова, кобра (или даже лев, корова, кобра).

Для сравнения заметим, что в случае обычного атрибутивного поиска возможно достаточно объективно судить о релевантности того или иного документа, вошедшего в выдачу, поскольку причиной выдачи нерелевантных документов являются погрешности в индексировании документов, проявляющиеся, например, во внесении в поисковый образ документа «лишних» слов (в результате явных ошибок, многозначности естественного языка и т. п.).

В каждом конкретном случае оценка пертинентности документа может быть более или менее объективно дана пользователем, ознакомившимся с этим документом, однако использование таких оценок для улучшения работы алгоритмов поиска «по аналогии» является иногда очень непростой задачей, так как связь между параметрами алгоритма и результатами выдачи далеко не всегда носит очевидный характер (особенно для алгоритмов, основанных на кластеризации).

Особенности алгоритмов поиска «по аналогии» для разных типов документов

Поиск научных публикаций, схожих с уже известными. Решение задачи кластеризации научных документов описано, например, в работе [15]. Для задания меры сходства на мно-

жестве документов было использовано расстояние Хемминга, подсчитываемое по формуле (2), где в качестве шкал использовались следующие атрибуты библиографического описания:

- авторы;
- ключевые слова;
- аннотация.

Так как сравнение аннотаций в явном виде (т. е. как текстовых строк), очевидно, бессмысленно, то они сравнивались как составные атрибуты на основании вхождения в их текст терминов из тезауруса соответствующей предметной области.

Как показало сравнительное тестирование ряда алгоритмов, наилучшие результаты для данной задачи показал FRiS-алгоритм [16].

При задании меры сходства был принят во внимание тот факт, что значения весовых коэффициентов в формуле (2) определяются предполагаемой апостериорной достоверностью данных соответствующей шкалы, и в определенных случаях один из коэффициентов может быть увеличен с пропорциональным уменьшением остальных.

Для определения весового коэффициента при каждом из атрибутов была проведена кластеризация выборок из базы данных «Сибирского математического журнала», в качестве критерия истинности применялся результат кластеризации, полученный с мерой, основанной на заранее заданных экспертами кодах классификатора MSC-2000.

Как показал эксперимент, наибольшее сходство с результатом кластеризации при мере, базирующейся на кодах классификатора, было достигнуто путем введения следующих продукционных правил.

1. Если каждый из документов d_1 и d_2 имеет более двух авторов и как минимум $^2/_3$ из числа авторов совпадают, то соответствующий весовой коэффициент при атрибуте «авторы» полагается равным единице.

2. Если каждый из документов d_1 и d_2 содержит более трех ключевых слов и как минимум $^3/_4$ этих слов совпадают, то соответствующий весовой коэффициент при атрибуте «ключевые слова» полагается равным единице.

3. Если каждый из документов d_1 и d_2 содержит более четырех терминов тезауруса в аннотации и как минимум $^3/_5$ этих терминов совпадают, то соответствующий весовой коэффициент при атрибуте «аннотация» полагается равным единице.

В противном же случае мы полагаем коэффициент при атрибуте «авторы» равным 0,2, а при атрибутах «ключевые слова» и «аннотация» равным 0,4.

Разумеется, для выборок, относящихся к другим предметным областям, количественные значения условий выбора коэффициентов и самих коэффициентов могут несколько отличаться от приведенных выше.

Обзор алгоритмов выявления дубликатов документов имеется в работе [17]

Объединение новостных сообщений в сюжеты. Данная задача отличается от предыдущей, в частности, тем, что новостные сообщения обычно структурированы намного слабее, чем научные публикации ввиду отсутствия у них основных атрибутов библиографического описания (авторы, аннотация, ключевые слова и т. п.).

Методы, применяемые для решения задачи объединения новостных сообщений в сюжеты, рассматриваются, например, в работе [18]. Отмечены два основных подхода: синтаксический и лексический.

Суть синтаксического подхода заключается в представлении документа в виде множества всевозможных последовательностей фиксированной длины k, состоящих из соседних слов (такие последовательности называются «шинглами»). Нетрудно видеть, что такие шинглы суть значения соответствующей номинальной шкалы. Два документа считаются похожими, если множества их шинглов существенно пересекаются. При этом если число совпадений слишком велико, документы считаются нечеткими дубликатами.

В рамках лексического подхода строится словарь (т. е. список различных слов) L коллекции документов, из которого исключены слова, встречающиеся в коллекции слишком редко и слишком часто (как правило, содержание документа наиболее адекватно отражают слова со средним значением частоты встречаемости). Далее для каждого документа формируется множество входящих в него различных слов U и определяется пересечение P этого списка с построенным словарем L. На основании близости таких списков можно судить о сходстве

документов. В частности, для выявления нечетких дубликатов списки слов P_i, соответствующие различным документам, упорядочиваются, и для каждого из них вычисляется хэш-функция. В случае совпадения хэш-функций документы объявляются нечеткими дубликатами.

Поиск художественных произведений. Эта задача по сравнению с предыдущими гораздо менее формализуема, поскольку свойствами художественных документов (книг, фильмов и т. п.), определяющими их сходство или различие, являются не только содержание, фамилии авторов и т. п., но и эстетическое впечатление, а в некоторых случаях – их идейная направленность. Поэтому основным приемом выявления сходства таких документа является экспертная оценка пользователей. Простейший прием такого рода встречается на некоторых сайтах, содержащих коллекции с описаниями художественных фильмов: пользователю предлагается указать известные ему фильмы, похожие на данный. Разумеется, такие оценки носят крайне субъективный характер.

Несколько более объективный подход используется в некоторых интернет-магазинах: в качестве одного из элементов описания книги или компакт-диска указывается список книг или дисков, которые пользователи обычно покупают вместе с данной книгой (диском).

Наконец, некоторые пользователи «Живого Журнала» и подобных ему ресурсов блогосферы не просто приводят списки своих любимых книг и фильмов, но и сопровождают их комментариями вида: «если Ваш список совпадает с моим более чем на треть, пожалуйста, сообщите свой список мне».

Заключение

Проведя неформальное обсуждение понятий «аналогии» и «сходства», мы изложили формализацию процедуры поиска объектов (прежде всего документов) «по аналогии». Эта процедура состоит, во-первых, в выделении основных составных частей, присущих данному типу объектов, во-вторых, в задании соответствующих шкал, содержащих множества возможных характеристик каждой из частей, в-третьих, в задании той или иной меры близости между объектами, и, наконец, в-четвертых, в выборе алгоритма принятия решения о сходстве объектов на основании вычисленных мер близости.

Еще раз подчеркнем: при поиске документов «по аналогии» оценка релевантности документа носит интенсиональный, а зачастую и весьма субъективный характер, поэтому качественная работа алгоритма может быть обеспечена в каждом конкретном случае лишь на основе тщательного анализа как предметной области поиска, так и целей пользователя, который предполагает воспользоваться результатами поиска.

Список литературы

1. *Философский* энциклопедический словарь. М.: Советская энциклопедия, 1983. С. 24.

2. *Новик И. Б., Уемов А. И.* Моделирование и аналогия // Материалистическая диалектика и методы естественных наук. М., 1968. С. 268–293.

3. *Толковый* словарь русского языка: В 4 т. / Под ред. Д. Н. Ушакова. М.: Гос. ин-т «Советская энциклопедия»; ОГИЗ; Гос. изд-во иностр. и нац. словарей, 1935–1940.

4. *Арский Ю. М., Гиляревский Р. С., Туров И. С., Черный А. И.* Инфосфера: информационные структуры, системы и процессы в науке и обществе. М.: ВИНИТИ, 1996.

5. *Солтон Дж.* Динамические библиотечно-информационные системы / Пер.с англ. М.: Мир, 1979.

6. *Новоселов М. М.* Сходство // Философский энциклопедический словарь. М.: Советская энциклопедия, 1983. С. 666.

7. *Ленин В. И.* Материализм и эмпириокритицизм // Ленин В. И. Полн. СОБР. соч. 5-е изд. М.: Политиздат, 1976. Т. 18.

8. *Ляпунов А. А.* О соотношении понятий материя, энергия и информация // Ляпунов А. А. Проблемы теоретической и прикладной кибернетики. Новосибирск: Наука, 1980. С. 320–323.

9. *Шрейдер Ю. А*. Равенство, сходство, порядок. М.: Наука, 1971.

10. *Тюхтин В. С.* Сходство // Большая советская энциклопедия. 3-е изд. М.: Советская энциклопедия, 1976. Т. 25. С. 123.

11. *Доброхотов А. Л.* Универсалии // Философский энциклопедический словарь. М.: Советская энциклопедия, 1983. С. 702–703.

12. *Воронин Ю. А.* Начала теории сходства. Новосибирск: Наука. Сиб. отд-ние, 1991.

13. *Пойа Д.* Математика и правдоподобные рассуждения / Пер. с англ. М.: Наука, 1975.

14. *Барсегян А. А., Куприянов М. В., Степаненко М. В., Холод И. И.* Методы и модели анализа данных: OLAP и Data Mining. СПб.: БХВ-Петербург, 2004.

15. *Барахнин В. Б., Нехаева В. А., Федотов А. М.* О задании меры сходства для кластеризации текстовых документов // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии. 2008. Т. 6, вып. 1. С. 3–9.

16. *Борисова И. А., Загоруйко Н. Г.* Функции конкурентного сходства в задаче таксономии // Материалы Всерос. конф. с междунар. участием «Знания – Онтологии – Теории» (ЗОНТ-07), Новосибирск, 14–16 сентября 2007 г. Т. 2. С. 67–76.

17. *Рубцов Д. Н., Барахнин В. Б.* Выявление дубликатов в разнородных библиографических источниках // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии. 2009. Т. 7, вып. 3. С. 86–93.

18. *Зеленков Ю. Г., Сегалович Ю. В.* Сравнительный анализ методов определения нечетких дубликатов для Web-документов // Тр. IX Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL’2007). Переславль-Залесский, 15–18 октября 2007 г. С. 166–174.

Материал поступил в редколлегию 14.10.2009

A. M. Fedotov, V. B. Barakhnin

PROBLEMS OF DOCUMENTS RETRIEVAL «IN ANALOGY»

Study of notion *analogy* and *similarity*, their formalization, and also procedure of information retrieval «in analogy» are resolved in the article.

Keywords: information retrieval, analogy, similarity, clustering.