Ю. В. Малинина*

СЕМАНТИЧЕСКАЯ СЕТЬ КАК ФОРМАЛЬНЫЙ МЕТОД ОПИСАНИЯ И ОБРАБОТКИ ТЕКСТОВ ПО ПРЕОБРАЗОВАНИЯМ ПРОГРАММ

ВВЕДЕНИЕ

В конце 50-х гг. в области программирования была поставлена задача повышения эффективности программ с помощью выполнения над ними в процессе трансляции ряда преобразований. Эту задачу принято называть задачей оптимизации программ, а преобразования, которые выполняются над программами и могут повысить их эффективность, — оптимизирующими преобразованиями. К настоящему времени теория оптимизации программ обогатилась фундаментальными результатами (развит аппарат схем программ, выделен богатый набор оптимизирующих преобразований, найдены достаточные условия корректности применения ряда преобразований и т.д.), а практика — примерами эффективно работающих оптимизирующих компиляторов. Исследования в этой области расширяются. А прогресс информационных технологий привел к тому, что общение исследователей стало глобальным, более интенсивным, а информационная насыщенность научных сообщений значительно возросла. Мировая система научной коммуникации предназначена для объединения целенаправленной деятельности большого числа ученых, проживающих в разных странах, говорящих на разных языках и работающих в одной тематической области.

Однако, по-прежнему, публикации являются основным источником информации и непременной составляющей научного исследования, одновременно представляют собой и цель, и средство (или же важную составляющую часть последних). В каком-то смысле исследование начинается с научного текста (изучение литературы, постановка задачи) и им же и заканчивается (опубликование результатов). Научный текст — это форма представления, формализации и обобщения научных знаний и одновременно — это способ аргументации и экспликации логического вывода и правдоподобных рассуждений.

* malin@iis.nsk.su

Работа выполнена при финансовой поддержке Министерства образования РФ (научная программа «Университеты России», грант № УР.04.01.201).

Сегодня многие ученые осознают, что развитие любого научного направления во многом зависит от привлечения в эту тематическую область новых научных сил. Для этого молодые люди, начинающие свой путь в науке, должны получить максимально возможное представление о состоянии тематики и о задачах, которые предстоит решать. Известно, что наиболее продуктивными являются те ученые, которые осознанно сделали свой выбор в направлении научных исследований.

В силу существования большого количества языков программирования, теоретические результаты в области оптимизации программ обычно формулируются в терминах абстрактных математических моделей программ, независимых от конкретных алгоритмических языков, причем число таких моделей исчисляется десятками [2, 5].

Это приводит к тому, что оптимизирующие преобразования описываются в терминах различных моделей, что затрудняет изучение различных оптимизирующих преобразований. Для решения этой проблемы наиболее ценными являются попытки собрать наиболее полные библиографии документов, относящихся к тематической области преобразования программ, на протяжении всей истории развития данного научного направления. Многие ученые понимают важность такой информационной работы, позволяющей сохранить для следующих поколений исследователей результаты их научной работы.

На сегодня текстовый поиск — это простейший способ доступа к текстовым данным, предназначенный для подбора материалов по выбранной теме. «Классическая» поисковая машина умеет найти по запросу из нескольких слов все документы, в которые данные слова входят, и предъявить их пользователю, что, кстати, может сделать и читатель печатного учебника, сравнив по предметному указателю, на каких страницах одновременно встречаются нужные ему термины.

Этой простой возможности при росте объемов текстовых баз становится совершенно недостаточно, и возникает задача предварительной обработки наборов текстовых документов с целью предоставления информации о них в более компактной и информационно насыщенной форме, одновременно сохраняя ссылки на исходные источники.

В середине 20-го века для упрощения поиска материала по определенной тематике использовались периодические реферативные издания, названия и тематическая направленность рубрик которых изменялась в соответствии с тенденциями развития отдельных областей науки. Причем в них впервые реферировались не только отдельные издания, но и публикации в периодических и продолжающихся изданиях. Для поиска документов в ре-

феративных журналах создавался справочный аппарат: авторский и тематический указатели в каждом номере и кумулятивный — за весь год.

Для электронных публикаций существует аналог — уникальный реферативный журнал, созданный институтом информации Гарфильда "Science Citation Index", ориентированный на поиск новых научных публикаций в мировой системе периодических и продолжающихся изданий по системе научных ссылок. Появление этого издания использует естественную исторически сложившуюся систему классификации научных работ по ссылкам автора на работы его предшественников в определенной тематической области. В этом издании отсутствует разбиение статей на тематические рубрики, авторы указателя предлагают пользователям подготовленные ими тематические кластеры связанных между собой публикаций по системе цитирования общих предшественников. Как размеры, так и наименования кластеров постоянно корректируются ими в соответствии с тенденциями в науке.

При сжатых сроках и при большом объеме публикующейся информации классификации научных работ по ссылкам автора на работы его предшественников в определенной тематической области зачастую недостаточно для оценки важности каждой публикации для ведущихся исследований и определения необходимости ее внимательного изучения. Для оценки требуется бóльшее количество параметров.

Поскольку даже беглый просмотр большого количества статей требует больших затрат времени, то для облегчения этой задачи обычно прибегают к различным технологиям автоматического свертывания документов

АВТОМАТИЗИРОВАННОЕ СВЕРТЫВАНИЕ

Рассмотрим одно из направлений интеллектуальной обработки научного текста — автоматизированное свертывание, к которому относятся работы по автоматическому реферированию, с точки зрения его применения для извлечения фрагментов определенного смысла.

Это направление занимает как бы промежуточное положение между минимальным уровнем свертывания — переводом и максимальным — индексированием. Фактически автоматическое реферирование по своему характеру очень специфично, поскольку сводится к экстрагированию (извлечению) из документов минимальных релевантных фрагментов, некоторая совокупность которых и образует широкий спектр вторичных документов — различные виды аннотаций, рефератов, реферативных аннотаций, самостоятельных фрагментов, конспектов и их синтезированных производ-

ных — реферативных указателей, дайджестов, реферативных обзоров, квазихрестоматий и т.д. Эти вторичные документы, являющиеся результатом аналитико-синтетической переработки первичного документного потока, рассчитаны на удовлетворение как частных (индивидуальных), так и типовых (потенциальных) информационных потребностей различных категорий специалистов науки, техники и производства [1].

Главное различие между средствами аннотирования состоит в том, что они, по существу, формируют краткое изложение или набор выдержек. Дополнительно рефераты различаются по функции и целевым группам пользователей [1]. Так, например, реферат может быть повествовательным, информативным или критическим.

Повествовательные рефераты формируются по классическому принципу извлечения информации: они предоставляют достаточный объем информации, чтобы создать у пользователя представление о соответствующих источниках с тем, чтобы их можно было отобрать для более внимательного прочтения.

Информативные рефераты заменяют собой текст, в основном они содержат основную или новую фактическую информацию в сокращенной форме.

Критические рефераты (или обзоры) сообщают не только суть информации, но и предлагают определенное мнение о ней. Критические рефераты обладают дополнительной ценностью по сравнению с оригиналом, поскольку предлагают выводы, которых нет в самом тексте.

Уже сегодня существуют действующие системы автоматического аннотирования текстовых документов. Прежде всего нужно сказать, что фактически во всех известных системах машинное аннотирование является экстрагированием — программа не «пересказывает» смысл текста, а просто извлекает из него те фрагменты, которые считает важными, и объединяет их в аннотацию [10]. Важность конкретного предложения определяется по различным параметрам, в частности, по так называемым маркерам важности (например, «в заключение нужно сказать, что...»), количеству содержательных слов в нем и т. д.

В наиболее развитых средствах аннотирования учитывается также зависимость предложений друг от друга с тем, чтобы не вносить в аннотацию обрывки, начинающиеся, например, со слов: «К тому же...», «В-третьих...» и т. п. [1].

Другой способ свертывания текста — это представление его в форме семантической сети, которая инвариантна к форме описания фактов с точностью до выбранной автором структуры рассуждения.

Понятие «семантической сети» известно в математике начиная с конца 50-х гг. Это помеченный граф, вершины которого используются для представления объектов (предметов, событий, состояний), а дуги — для представления связей (отношений между объектами). Такая структура хорошо изучена с точки зрения математики и служит удобным средством представления знаний для дальнейшего анализа.

Сжатие информации при переходе от лексического к семантическому описанию документов приводит к ее *обобщению*, что эквивалентно получению некоторого *знания*. Ведь возможность более сжатого описания данных есть следствие скрытых в этих данных закономерностей. Сжатие информации как раз и сводится к выявлению этих закономерностей, выражающих наши знания о структуре данных.

Семантическая сеть — это множество понятий (слов и словосочетаний), связанных между собой. В семантическую сеть включаются наиболее часто встречающиеся слова текста, которые несут основную смысловую нагрузку. Для каждого понятия формируется набор ассоциативных (смысловых) связей, т.е. список других понятий, в сочетании с которыми оно встречалось в предложениях текста. При этом считается, что чем чаще встречаются вместе два понятия в предложениях текста, тем выше вероятность того, что они связаны по смыслу.

Статистическая информация об отдельных лексических единицах легко извлекается из текста, и есть все основания полагать, что она адекватно отражает его содержание в целом. Косвенное подтверждение этому можно найти в нейропсихологических исследованиях, которые установили, что анализ печатного текста, опираясь на зрительное пространственное (а не на линейное слуховое) восприятие, реализуется преимущественно правым полушарием мозга, использующим ассоциативную статистическую модель [3, 6]. Логический «левополушарный» анализ, моделированием которого, по сути, занимается формальная лингвистика, необходим лишь в отдельных «трудных» местах текста, несущих новую информацию и требующих детального осмысления.

АЛГОРИТМ РЕАЛИЗАЦИИ

Общая схема обработки текстов инвариантна по отношению к выбору естественного языка. Независимо от того, на каком языке написан исходный текст, его анализ проходит одни и те же стадии. Первые две стадии (разбиение текста на отдельные предложения и на слова) практически оди-

наковы для большинства естественных языков. Единственное, где могут проявиться специфичные для выбранного языка черты, — это обработка сокращений слов и знаков препинания (точнее, определение того, какие из знаков препинания являются концом предложения, а какие — нет). Последующие стадии: морфологический анализ; морфемный анализ; синтаксический анализ, напротив, сильно зависят от выбранного естественного языка. Последняя стадия (семантический анализ) также мало зависит от выбранного языка, но это проявляется только в общих подходах к проведению анализа [9].

В общем виде семантическую сеть можно представить как взвешенный граф, в вершинах которого находятся знаки, а ребра с весами отражают связи знаков между собой. Рассмотрим ассоциативную семантическую сеть. Для определения понятия близости двух знаков в таких сетях требуется для каждого из двух изучаемых знаков составить список часто встречающихся рядом с ним знаков. Таким образом, для каждого из двух знаков мы получим список «ассоциированных» с ним знаков. Сравнивая эти списки, можно сделать количественные выводы о том, насколько близки исходные два знака.

Текст будем рассматривать как множество соответствующих основ слов, входящих в него. Две основы слова объявим близкими, если они встречаются в тексте на расстоянии k слов. Знаки препинания и другие специальные символы и стоп-слова, встречающиеся в тексте, игнорируем. Степень близости двух слов — это число случаев, когда два слова, соответствующих данному знаку, встретились рядом. Последовательность обработки текста будет следующая.

1. Построить упорядоченный по очередности вхождения списка слов текст, с учетом пропуска так называемых стоп-слов. Это наиболее употребительные в данном языке слова, удаление которых не повлияет на качество идентификации (более того, может его улучшить). Например, в английском языке имеются стандартные доступные в Internet списки стоп-слов (несколько сотен слов), включающие в свой состав артикли, союзы, модальные и вспомогательные глаголы, числительные, местоимения.
2. Выделить основы слов (stemming), с помощью подходов, описанных в [2]. Это позволяет все однокоренные слова заменить корнем и не различать, например, слова task и tasks.
3. Исходя из того, как часто слова появляются рядом, для каждого слова составить списки близких к нему слов.

4. Сравнивая списки любых двух знаков, определить меру близости этих знаков.
5. По полученной информации построить взвешенный граф со знаками в качестве вершин и взвешенными ребрами в качестве связей между знаками.

Таким образом, для фразы «Неопределенность зависимостей между индексированными переменными» получаются следующие знаки: (определ, зависим, индекс, перемен). При k = 1 со знаком «зависим» будут рядом знаки (определ, индекс) с весами (1, 1) соответственно.

Описанный подход содержит упрощенный способ формирования семантической сети по тексту. Предполагается, что в дальнейшем он будет расширен вскрытием всех типов связей и взаимоотношений между понятиями. Основные виды таких отношений — это гипонимия (род—вид), соподчинение на одном уровне — парциация (часть—целое), ассоциация (локализация объекта, его назначения). В каждой системе знания любой объект в каком-то отношении нередко является и признаком, а почти каждый признак в другом отношении выступает как объект.

ЗАКЛЮЧЕНИЕ

Выше в краткой форме представлено состояние работы и сформулированы некоторые подходы к решению. Однако рамки задачи автоматизированного свертывания документного потока должны рассматриваться значительно шире. Речь идет о создании системы автоматической обработки потока публикаций с целью максимального раскрытия и использования его ресурсов для решения задач развития науки.

Не секрет, что в системе информационных коммуникаций наблюдается постоянное недоиспользование накопленных обществом знаний со всеми вытекающими отсюда негативными последствиями. Причина этого, прежде всего, в несовершенстве средств поиска информации (несмотря на широкое внедрение в эту сферу средств компьютерной техники) и методов аналитико-синтетической переработки первичного документального потока. Специалисту в действительности нужны не документы, а информация — факты, концепции и др. Однако информации очень много вообще, но крайне мало — в частности.

Такое положение обусловлено диалектическим противоречием между избыточностью конкретного документа за счет его многоаспектности и недостаточностью документного потока в целом за счет явления рассеяния.

Работы в области информационного анализа/синтеза и призваны, в известных рамках, снять это противоречие. Их конечная цель — максимальное использование когнитивных (познавательных) возможностей первичного документа за счет машинного «разбиения» его на самостоятельные минимальные релевантные фрагменты, утилизируемые затем в гипотетической пока еще базе знаний, обращение к которой позволило бы в значительной степени снизить необходимость использования первичного потока.

Предложенная работа является лишь предварительным этапом в реализации идеи компьютерного свертывания, которая должна постепенно трансформироваться в серию работ, направленных на получение репрезентативных результатов.

СПИСОК ЛИТЕРАТУРЫ

1. **Блюменау Д.И.** О некоторых направлениях формализации инфопроцессов // Проблемы инфовзаимодействия. — Новосибирск, 1993. — С. 206–223.
2. **Евстигнеев В. А. , Мирзуитова И. Л.** Анализ циклов: выбор кандидатов на распараллеливание. — Новосибирск, 1999. — (Препр. / ИСИ СО РАН; № 58).
3. **Евстигнеев В.А.** О некоторых формах промежуточного представления программ // Оптимизация и конструирование программ. — Новосибирск, 1993. — С. 52–59.
4. **Ермаков А.Е., Плешко В.В.** Семантическая сеть текста в задачах аналитика // Информатизация и информационная безопасность правоохранительных органов: Тр. XI Междунар. научной конф. — Москва, 2002. — С. 343–347.
5. **Ермолаев Д.С** Компьютерный морфологический разбор слов русского языка. — 2001 — `http://www.codenet.ru/progr/alg/morf.php`
6. **Иванко Е., Перевалов Д.** Использование ассоциативных семантических сетей для классификации звукозаписей // Компьютерная лингвистика и интеллектуальные технологии: Тр. Междунар. конф. «Диалог'2004». — М.: Наука, 2004.
7. **Касьянов В.Н**. Оптимизирующие преобразования программ. — М.: Наука, 1988.
8. **Киселев С.Л., Ермаков А.Е., Плешко В.В.** Поиск фактов в тексте естественного языка на основе сетевых описаний // Компьютерная лингвистика и интеллектуальные технологии: Тр. Междунар. конф. «Диалог'2004». — М.: Наука, 2004.
9. **Коваленко А.** «СТЕММЕР» морфологический анализ для небольших поисковых систем. —2002. — `http://samag.ru/img/uploaded/samag14649-0.pdf`
10. **Селезнев К**. Обработка текстов на естественном языке // Открытые системы. — 2003. — № 12.
11. **Хан У., Мани И**. Системы автоматического реферирования // Открытые системы. — 2003. — № 12.