МЕЖДУНАРОДНЫЙ СОЛОМОНОВ УНИВЕРСИТЕТ

ЛАНДЭ Дмитрий Владимирович

ТЕОРИЯ ИНФОРМАЦИОННОГО ПОИСКА

Методическое пособие
для студентов кафедры компьютерных наук

(краткое содержание магистерского курса лекций)

Киев-2006

Теория информационного поиска / Методическое пособие/
Краткое содержание магистерского курса лекций /

© Д.В. Ландэ, к.т.н., доцент МСУ

Аннотация

На современном этапе специалистам в области информационных технологий необходимо иметь представление, теоретические и практические навыки роботы с информационно-поисковыми системами, в частности, с поисковыми системами в сети Интернет.

Именно теоретическим вопросам организации поиска в информационных ресурсах сети Интернет, навигации в этой сети посвящен данный краткий лекционный курс, рассчитанный на 28 часов. На занятиях рассматриваются вопросы, относящиеся к информационной структуре Web-пространства, полнотекстовым информационно-поисковым системам, их алгоритмическому и лингвистическому обеспечению, возможностям ранжирования, аналитического обобщения результатов поиска, общим закономерностям современного информационного пространства.

Курс лекций расчитан на студентов магистратуры, специализирующихся в информационных технологиях, и соответственно, имеющих достаточную подготовку в таких областях математики, как высшая алгебра, дифференциальные уравнения и теория вероятностей.

Лекции данного курса подкрепляются практическими занятиями (28 часов), в рамках которых студенты получают навыки работы с информационно-поисковыми системами в Интернет, системой контент-мониторинга InfoStream (© *ElVisti*), популярными пакетами статистической обработки данных.

СОДЕРЖАНИЕ

ЛЕКЦИЯ 1. ПРЕДМЕТ И СОДЕРЖАНИЕ ДИСЦИПЛИНЫ. ОБЩАЯ ИНФОРМАЦИЯ ОБ ИНТЕРНЕТ, ГИПЕРТЕКСТЕ, WEB-ПРОСТРАНСТВЕ

1) Общая информация об Інтернет

Интернет - глобальная информационная сеть, части которой логически взаимосвязаны друг с другом посредством единого адресного пространства, основанного на протоколе TCP/IP. Интернет состоит из множества взаимосвязанных компьютерных сетей и обеспечивает удаленный доступ к компьютерам, электронной почте, доскам объявлений, базам данных и дискуссионным группам. *(Глоссарий.ru)*

- История (> 30 лет, ARPANET, США).
- Объемы (> 20 млрд. документов, > 90 млн. сайтов – авг. 2006 г.).
- Статистика роста (www.netcraft.com).
- Темпы роста (7-10 млн. документов в день).

Благодаря чему именно Интернет получил такое развитие?

- Высокая технологичность, надежность и устойчивость.
- Открытость протоколов.
- Поддержка пользователями и производителями ПО.
- Способность к саморазвитию, саморасширению.
- Постоянное снижение затрат абонентов на работу в Интернет.
- Де-факто новый вид интерактивного СМИ.

Буква *S* технологического процесса – до 1990 гг. –развитие ресурсов и достижение насыщения (среда обмена данными), после 1990 г. – новый этап бурного развития, связанный с появлением Web-технологий.

2) Гипертекст и Интернет

Гипертекст - принцип организации информационных массивов, при котором отдельные информационные элементы связаны между собой ассоциативными отношениями, обеспечивающими быстрый поиск необходимой информации и/или просмотр взаимосвязанных данных. *(Глоссарий.ru)*

Гипертекст – форма организации взаимосвязи между отдельными фрагментами текста.

Информация из истории гипертекстовых технологий:

В 1945 году Ванневер Буш (Vannevar Bush) создал первую фотоэлектрическую память и приспособление Memex (memory extension), представляющее собой справочник, реализованный путем гиперссылок в пределах документа. Тед Нельсон (Ted Nelson) в 1965 году ввел термин "гипертекст" и создал гипертекстовую систему Xanadu с двухсторонними гиперсвязями.

В 1980 году Тим Бернерс-Ли (Berners-Lee), консультант CERN (Европейская организация ядерных исследований) написал программу, позволяющую создавать и просматривать гипертекст, реализующая двунаправленные связи между документами в коллекции [69].

В 1990 году для поддержки документации, циркулирующей в CERN Бернерс-Ли начал работу над графическим интерфейсом пользователя (GUI) для гипертекста. Эта программа была названа "WorldWideWeb". К 1992 году уже были созданы такие GUI как Erwise и Viola.

Браузеры – программы визуализации гипертекста в среде WWW.

В феврале 1993 года Марк Андрессен (Mark Andressen) из NCSA (Национальный Центр Суперкомпьютерных приложений США, www.ncsa.uiuc.edu) закончил начальную версию программы визуализации гипертекста Mosaic для популярного графического интерфейса Xwindow System под UNIX. Одновременно CERN развивал и улучшал HTML - язык гипертекстовой разметки текстов, и HTTP - протокол передачи гипертекста, а также сервер обработки гипертекстовых документов - CERN HTTPD.

Протокол передачи гипертекстовых данных HTTP.

HTTP (HyperText Transfer Protocol) - это протокол передачи Web-страниц по сети Интернет. Изначально протокол HTTP использовался исключительно для передачи HTML - документов, но в настоящее время посредством HTTP можно передавать любую информацию, в том числе картинки, звук, видео, а также просто абстрактные файлы.

HTML - основной язык гипертекстовой разметки в Интернет.

Преимущества и недостатки:

- простота;
- ориентация на эффективную визуализацию, оформление;
- нет ориентации на автоматическую обработку;
- нет однотипности представления контента, что затрудняет поиск.

3) Модель Web-пространства

В ноябре 1999 года один из руководителей института поиска и анализа текстов, входящего в исследовательское подразделение IBM, Андрей Брёдер (Andrei Broder) и его соавторы из компаний AltaVista, IBM и Compaq математически описали "карту" ресурсов и гиперсвязей Web.

Проследив с помощью поискового механизма AltaVista свыше 200 млн. Web-страниц и несколько миллиардов ссылок, размещенных на этих страницах, ученые пришли к выводу о структуре Web-пространства как ориентированного графа, в котором вершины соответствуют Web-страницам, а ребра – соединяющим эти страницы гиперссылкам. В рамках этой модели задача анализа структуры связей между отдельными Web-страницами было обнаружено:

- центральное ядро (28% Web-страниц) - компоненты сильной связности (SCC).
- 22% Web-страниц - это "отправные Web-страницы" (IN). Они содержат гиперссылки, которые в конечном счете ведут к ядру, но из ядра к ним попасть нельзя.
- столько же - 22% - "оконечных Web-страниц" (OUT), к которым можно прийти по ссылкам из ядра, но нельзя вернуться назад.
- 22% Web-страниц - отростки - полностью изолированы от центрального ядра: это либо "мысы", связанные гиперссылками со страницами любой другой категории, либо "перешейки", соединяющие две Web-страницы, не входящие в ядро.

В модели учтены и "острова", которые вообще не пересекаются с остальными ресурсами Internet. Единственный способ обнаружить ресурсы этой группы - знать адрес.

4) Статическая и динамическая составляющие Web-пространства:

- информация долгосрочного характера.
- обновляемая информация.

Модель Бартона-Кеблера учитывает обе составляющие:

$$m(t) = 1 - ae^{-T} - be^{-2T},$$

где m(t) – доля полезной информации.

5) Web-порталы

Современные интегрированные Интернет-ресурсы, реализующие функции:

- Информационный сервис (поиск и получение информации).
- Бизнес-функции.
- Инструментарий пользователя, помогающий ему создавать свой контент.
- Сервис общения.

6) «Скрытый» Web

Совокупность Web-ресурсов, не видимая «глобальными» поисковыми системами, в том числе динамически формируемые Web-страницы и документы из баз данных.

Некорректность расчета объемов «островов» по Бредеру ввиду предопределенного списка Web-ресурсов из базы данных системы AltaVista.

Каталоги «скрытого» Web:

- Complete Planet (http://www.completeplanet.com/) – каталог, содержащий актуальный список свыше 20000 ресурсов "скрытого" Web.
- Librarians' Index to the Internet (http://lii.org/) – каталог, содержащий свыше 14000 Internet-ресурсов. LII также включает ссылки на "скрытые" в Web-пространстве базы данных. Есть ссылка "and databases" (добавить базу данных).
- FindLaw (http://www.findlaw.com/) –каталог правовых ресурсов, содержащий аннотированный список свободно доступных баз данных нормативно-правовых документов, для которых данный ресурс является «точкой входа».
- InfoMine (http://infomine.ucr.edu) - ресурс, содержащий ссылки на 120000 документов, представленных в 9 аннотированных баз данных.

Попытки интеграции доступа к объектам «скрытого» Web – системы поиска в «скрытом» Web.

7) Основные проблемы, связанные с развитием Интернет-контента

- Прогресс в области производства информации ведет к снижению уровня информированности.
- Интенсивность роста объемов шумовой информации во много раз превышает интенсивность роста объемов полезной информации.

Литература

Публикации Д.В. Ландэ по тематике лекции:

1. Ландэ Д.В. Поиск знаний в Internet. –М.: Диалектика-Вильямс, 2005. (http://poiskbook.kiev.ua)
2. Ландэ Д.В. Основы интеграции информационных потоков.– К.:Инжиниринг, 2006. (http://dwl.visti.net/art/monogr-osnov/spusk3.pdf)
3. Ландэ Д.В. Навигация в Сети: каталоги – поисковики – порталы. // InternetUA. -2000. - № 1. - С. 43-47.(http://dwl.visti.net/art/nav/)
4. Ландэ Д.В. Затерянный вэб. // "Телеком" № 1-2, 2005 (http://dwl.visti.net/art/zw/)
5. Ландэ Д.В. Интернет для людей. // Бизнес-регистр. –№5(11). – 2002 (http://dwl.kiev.ua/art/obzor/)

Публикации других авторов:

1. В. Гусев. Освоение Internet. Самоучитель. – М: Вильямс, 2003.
2. А. Левин. Интернет - это очень просто! – СПб: Питер, 2004.
3. Чурсин Н.Н. Популярная информатика. -К.: Техника, 1982. -158 с.
4. Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, Janet Wiener. Graph structure in the Web. Proceedings of the 9th international World Wide Web conference on Computer networks: the international journal of computer and telecommunications networking. Amsterdam, The Netherlands, 2000. - P. 309-320. (*http://www.almaden.ibm.com/cs/k53/www9.final/*)

5. The Deep Web: Surfacing Hidden Value, 2000 BrightPlanet.com LLC, 35 p. (http://www.dad.be/library/pdf/BrightPlanet.pdf)

ЛЕКЦИЯ 2. ОБЩИЕ СВЕДЕНИЯ ОБ ИНФОРМАЦИОННО-ПОИСКОВЫХ СИСТЕМАХ

1) История информационно-поисковых систем (Fulltext Retrieval System)

Информационно-поисковая система - система, выполняющая функции: - хранения больших объемов информации; - быстрого поиска требуемой информации; - добавления, удаления и изменения хранимой информации; - вывода информации в удобном для человека виде. *(Глоссарий.ru)*

Первые полнотекстовые информационно-поисковые системы появились в начале компьютерной эры. Назначением этих систем был поиск в библиотечных каталогах, архивах, массивах документов, таких как, статьи, нормативные акты, рефераты, тексты брошюр, диссертаций, монографий. В начале информационно-поисковые системы (ИПС) использовались преимущественно в библиотечном деле и в системах научно-технической информации.

1965 – 1970:
- Dialog, MARK, STAIRS

1990 – 1995:
- Z39.50, Galileo, WAIS

1995 – 2006:
- RetrievalWare, Autonomy, AltaVista, Яндекс, Google…

2) Сетевые ИПС

Сегодня миллионам пользователей Интернет известны такие информационно-поисковые системы, как Google, Yahoo, AltaVista, AllTheWeb, MSN … Яndex, Рамблер, которые охватывает миллиарды Web-документов.

В отличие от реляционных СУБД, у систем полнотекстового поиска не существует стандартизированного языка запросов. У каждой системы этого типа существует свой способ задания критериев поиска. Очень часто языки запросов поисковых систем приближены к SQL, однако каждой из них присущ ряд индивидуальных особенностей, связанных с такими моментами, как:

- интерпретация операций, зависящих от порядка расположения слов в тексте (операций контекстной близости слов и др.);
- реализация вычисления близости найденных документов запросам (релевантности) для представления результатов поиска;
- применение нестандартных функций, требующих, например, использования методов искусственного интеллекта (нахождение документов по принципу подобия, построение дайджестов из фрагментов документов, сниппетов и др.)

В различных полнотекстовых ИПС различаются архитектуры, структуры данных, алгоритмы их обработки, методы организации поиска.

2) Характеристики ИПС:

- Полнота
- Релевантность

Понятие пертинентности как характеристики информационно-поисковой системы, означающее соответствие полученной информации информационной потребности.

Таблица оценки качества ИПС в TREC (РОМИП):

Документы	Выданные	Не выданные
Релевантные	a	c
Не релевантные	b	d

Коэффициент полноты:
$p = a / (a + c)$
Коэффициент точности:
$n = a / (a + b)$
Коэффициент осадков:
$q = b / (a + b)$
Коэффициент специфичности:
$k = d / (b + d)$

11-точечный график полноты/точности TREC (РОМИП)

11-точечный график полноты/точности отражает изменение точности в зависимости от требований к полноте и дает более полную информацию, чем единая метрика в виде одной цифры. По оси абсцисс на графике откладываются значения полноты, по оси ординат – значение точности при условии, что рассматривается начальный отрезок результатов запроса, на котором достигается заданный уровень полноты. Для запроса, которого известно n релевантных документов, полнота может принимать дискретные значения 0, 1/n, 2/n, ... , 1.

Для того, чтобы можно было получать единый график полноты/точности для множества запросов

1. рассматриваются фиксированные значения полноты 0.0, 0.1, 0.2, ... , 1.0 (всего 11 значений);
2. используется специальная процедура интерполяции точности для данных фиксированных значений полноты;
3. для множества запросов производится усреднение точности для заданных уровней полноты.

Интерполированное значение точности равно максимальному значению точности при уровне полноты большем или равным заданному.

Подробно описывается процедура построения 11-точечного графика, а также пример построения графика.

3) Технологические характеристики:

- скорость обработки запросов;
- полнота охвата ресурсов;
- вероятность получения ответа от системы;
- нахождение документов, подобных найденным;
- возможность уточнения запросов;
- возможность подключения переводчиков и т.д.

4) Недостатки традиционных информационно-поисковых систем

- недостаточная оперативность;
- зависимость от выбора источников;

- слабые поисковые возможности;
- отсутствие средств уведомления о нахождении новой информации;
- недостаточная защита данных;
- слабо развитые средства обобщения данных.

Литература

Публикации Д.В. Ландэ по тематике лекции:

1. Ландэ Д.В. Поиск знаний в Internet. –М.: Диалектика-Вильямс, 2005. (http://poiskbook.kiev.ua)
2. Ландэ Д.В. Поисковые системы: поле боя – семантика. // "Телеком" № 4, 2004 (http://dwl.visti.net/art/abstr-st/index.html)
3. Ландэ Д.В. Искать и не сдаваться. // CHIP/Украина. – 2004. – №5. - С. 84-87 (http://dwl.visti.net/art/sart/)

Публикации других авторов:

1. Попов А. Поиск в Интернете - внутри и снаружи // Intrnet. – 1998. № 2. (http://www.citforum.ru/pp/search_03.shtml).
2. Гусев В.С. Поиск в Internet: Самоучитель. СПб: Диалектика /Вильямс, 2004. - 336 с.
3. Program to evaluate TREC results using SMART evaluation procedures. Documentation. (http://wwwnlpir. nist.gov/projects/trecvid/trecvid.tools/trec_eval/README).
4. Труды четвертого российского семинара РОМИП'2006. (Суздаль, 19 октября 2006г.) Санкт-Петербург: НУ ЦСИ, 2006, 274 с.. (http://romip.narod.ru/romip2006/index.html)
5. И. Кураленок, И. Некрестьянов. Оценка систем текстового поиска. / Программирование. - 2002. - № 28(4). -С. 226-242.
6. Ricardo Baeza-Yates and Berthier Ribeiro-Neto. Modern Information Retrieval. ACM Press, 1999. – 513 p.
7. Сегалович И.В. Как работают поисковые системы. // Мир Internet. – 2002. -№ 10. (http://www.dialog-21.ru/direction_fulltext.asp?dir_id=15539).
8. RFC 1625 - WAIS over Z39.50-1988. Network Working Group. Request for Comments: 1625. M. St. Pierre, J.Fullton, K.Gamiel, J.Goldman, B.Kahle, J.Kunze, H.Morris, F.Schiettecatte, 1994. (http://www.faqs.org/rfcs/rfc1625.html)

ЛЕКЦИЯ 3. ОСНОВНЫЕ МОДЕЛИ ПОИСКА

1) Булева модель поиска

Базируется на теории множеств и математической логике.

Каждый запрос – логическое выражение, связываемое операторами AND, OR, NOT.

Архитектура ИПС, базирующихся на булевой модели.

Пример – организация наборов данных в ИПС STAIRS (IBM).

Состав таблиц ИПС с инвертированными списками:

- текстовая;
- указатели на тексты;
- словарь уникальных слов;
- инверсная, содержащая списки номеров документов, соответствующих определенным словам.

Описание процесса поиска информации в ИПС с инвертированными списками:

- обращение к словарю уникальных слов;
- обращение к инверсной таблице;
- обращение к указателям на тексты;
- обращение к текстовой таблице.

2) Векторно-пространственная модель поиска

Классическая алгебраическая модель. Документ описывается вектором в некотором евклидовом пространстве. Каждому терму сопоставляется вес, характеризующийся частотой, местоположением, тематикой и т.п.

Запрос – также вектор в евклидовом пространстве. Близость запроса документу – скалярное произведение.

*Подход к взвешиванию термов/документов – TF*IDF:*

TF – частота появления терма в документе;

IDF – величина, обратная количеству документов массива, которые содержат данный терм.

Векторно-пространственная модель обеспечивает:

- обработку запросов без логических ограничений их длины;
- простоту реализации режима поска подобных документов;
- сохранение результатов поиска с возможностью выполнения уточняющего поиска.

3) Вероятностная модель поиска

Базируется на теоретических основах байесовой условной вероятности.

Основной подход – вероятностная оценка весов терминов в документах.

Функционирование модели базируется как на экспертных оценках пользователей, которые признают документ релевантным/нерелевантным, так и на априорных оценках вероятности того, что документ является релевантным исходя из состава входящих в него терминов.

Первоначально в вероятностной модели использовалось упрощение, предполагающее независимость вхождения в документ любой пары термов («наивный» байесовский подход).

В случае применения экспертных оценок процесс поиска – итерационный. На каждом шагу итерации, благодаря режиму обратной связи, определяется множество документов, отмеченных пользователем как удовлетворяющих его информационным потребностям.

Модель широко используется для решения вспомогательных задач: определения тональности сообщений, выявления спама и т.д.

4) Байесовский подход к решению проблемы спама

В методе Байеса подразумевается использование оценочной базы — двух корпусов электронных писем, один из которых составлен из спама, а другой — из обычных писем. Для каждого из корпусов подсчитывается частота использования каждого слова, после чего вычисляется весовая оценка (от 0 до 1), характеризующая условную вероятность того, что сообщение с этим словом является спамом. Значения весов, близкие к ½, не учитываются при интегрированном расчете, поэтому слова с такими весами игнорируются и удаляются из словарей.

В соответствии с методом, предложенным Полом Грэмом (*Paul Graham*), если сообщение содержит n слов с весовыми оценками $w_1 ... w_n$, то оценка условной вероятности того, что письмо окажется спамом, основанная на данных из оценочных корпусов, вычисляется по формуле:

$Spm = \prod w_i / (\prod w_i + \prod (1-w_i))$.

Приведенная формула обосновывается следующим соображением. Предполагается, что S – событие, заключающееся в том, что письмо – спам, A – событие, заключающееся в том, что письмо содержит слово t. Тогда, в соответствии с формулой Байеса, справедливо:

$$P(S \mid A) = \frac{P(A|S)P(S)}{P(A|S)P(S)+P(A|\bar{S})P(\bar{S})},$$

Если изначально не известно, является письмо спамом или нет, исходя из опыта предполагается, что $P(\bar{S}) = \lambda P(S)$, из чего следует:

$$P(S \mid A) = \frac{P(A|S)}{P(A|S)+\lambda P(A|\bar{S})}$$

Далее формула обобщается следующим образом. Предполагается, что A_1 и A_2 – это события, заключающиеся в том, что письмо содержит слова t_1 и t_2. При этом вводится допущение, что эти события независимы (поэтому метод называется «наивным» байесовским). Условная вероятность того, что письмо, содержащее оба слова (t_1 и t_2) является спамом, равна:

$$P(S \mid A_1 \,\&\, A_2) = \frac{P(A_1|S)P(A_2|S)}{P(A_1|S)P(A_2|S)+\lambda P(A_1|\overline{S})P(A_2|\overline{S})} = \frac{p(t_1)p(t_2)}{p(t_1)p(t_2)+\lambda(1-p(t_1))(1-p(t_2))}$$

Обобщением формулы на случай произвольного количества слов и λ=1 и является формула П. Грэма.

Следует отметить, что широкое применяемое в находит именно значение λ=1. Хотя это несколько упрощает вычисления, но серьезно искажает действительность и снижает качество. На практике на основе словарей, которые постоянно модифицируются, для каждого сообщения рассчитывается значение *Spm*. Если оно больше некоторого порогового, то сообщение считается спамом.

5) Недостатки рассмотренных моделей:

- Булева модель – невысокая эффективность поиска, жесткий набор операторов, невозможность ранжирования.
- Векторно-пространственная модель связана с расчетом массивов высокой размерности, малопригодна для обработки больших массивов данных.
- Вероятностная модель характеризуется низкой вычислительной масштабируемостью, необходимостью постоянного обучения системы.

Приведенные классические модели изначально предполагали рассмотрение документов как множества отдельных слов, не зависящих друг от друга. Такая упрощающая концепция имеет название «Bag of Words». В реальных системах это упрощение преодолевается, например, расширенная булева модель учитывает контекстную близость (операторы NEAR, ADJ во многих известных системах). Системы базирующиеся на вероятностной модели учитывают вхождение словосочетаний и связи отдельных терминов, хотя большинство из известных систем борьбы со спамом, построенные на вероятностной модели все-таки базируются на упрощенном подходе независимости отдельных слов.

Кроме представленных существуют и другие методы поиска, например, семантические, в рамках которых делаются попытки организации смыслового поиска за счет анализа грамматики текста, использования баз знаний, тезаурусов, онтологий, реализующих семантические связи между отдельными словами и их группами. Такие подходы пока остаются очень затратными, область их применения – профессиональные аналитические системы.

Литература

Публикации Д.В. Ландэ по тематике лекции:

1. Ландэ Д.В. Архитектура ИПС в среде реляционной СУБД. // Інформатизація та нові технології. – К., 1994. - № 1-2. - С. 28-30. (http://dwl.visti.net/art/zzz/rsdb1.html, http://dwl.visti.net/art/zzz/rsdb2.html, http://dwl.visti.net/art/zzz/rsdb3.html)
2. Ландэ Д.В. Основы интеграции информационных потоков. – К.: Инжиниринг, 2006. – 240 с. (http://dwl.visti.net/art/monogr-osnov/spusk3.pdf)
3. Ландэ Д.В, Зубок В.Ю. Информационно-поисковый сервер InfoReS для работы в среде WWW. // Компьютеры плюс программы. – 1996. -№ 5. - С. 65-69. (http://dwl.visti.net/art/cgi/)
4. Ландэ Д.В. Поиск знаний в Internet. –М.: Диалектика-Вильямс, 2005. (http://poiskbook.kiev.ua)

Публикации других авторов:

1. Ricardo Baeza-Yates and Berthier Ribeiro-Neto. Modern Information Retrieval. ACM Press, 1999. – 513 p.

2. Сегалович И.В. Как работают поисковые системы. // Мир Internet. – 2002. -№ 10. (http://www.dialog-21.ru/direction_fulltext.asp?dir_id=15539).
3. А. В. Аграновский, Р.Э. Арутюнян. Индексация массивов документов // Мир ПК, - № 06. -2003 (http://old.osp.ru/pcworld/2003/06/049.htm)
4. Salton G., Fox E., and Wu H. Extended Boolean information retrieval. Communications of the ACM. – 2001. - Vol. 26. - №. 4. - P. 35-43.
5. Salton G, Wong A, Yang. C. A Vector Space Model for Automatic Indexing. // Communications of the ACM, 18(11):613-620, 1975.

ЛЕКЦИЯ 4. ИНФОРМАЦИОННО-ПОИСКОВЫЕ ЯЗЫКИ. ИНТЕРФЕЙСЫ ПОЛЬЗОВАТЕЛЕЙ ИПС

1) Лингвистическое обеспечение ИПС

- Информационно-поисковые языки (ИПЯ), то есть языки, на которых обращаются пользователи к системе.
- Языки представления данных в ИПС.
- Естественные языки и языки разметки, на которых представлены документы-первоисточники.

> Информационно-поисковый язык - формализованный искусственный язык, предназначенный для индексирования документов, информационных запросов и описания фактов с целью последующего хранения и поиска. *(Глоссарий.ru)*

2) Возможности ИПЯ: поиск по словам, усечениям и словоформам

Все поисковые системы обеспечивают поиск хотя бы по ***одному слову***.

Некоторые системы рассматривают все слова запроса как ***правые усечения***.

У некоторых такая возможность не реализована, например у Google, Alltheweb.

Поиск по словоформам является результатом серьезного лингвистического анализа и реализован в русскоязычных системах Апорт, Яndex и Рамблер, а также в укранской системе МЕТА.

Большинство современных систем способно реализовывать ***контекстный поиск заключенной в кавычки фразы*** (Google, Alltheweb, AltaVista, Lycos и др.). Такая способность - это реализация неявно указанных с помощью кавычек операторов контекстной близости.

3) Возможности ИПЯ: операторы

В большинстве современных систем реализованы булевы операторы AND, OR и NOT, а также работа со скобками. В свое время функции контекстной близости получили наибольшее развитие в системе Lycos, где были реализованы с помощью четырех операторов: ADJ, NEAR, FAR и BEFORE.

Можно отметить, что у самой популярной в мире системы Google - самый лаконичный набор логических операторов - "+", OR и "-".

Слова запроса, заключенного в двойные кавычки, ищутся в документах в том порядке и в тех формах, в которых они встретились в запросе.

При построении запросов может быть реализована возможность объединения слов в группы, которые будут аргументами некоторого оператора. Такие группы заключаются в скобки. Использование скобок позволяет создавать вложенные запросы, а также изменять приоритеты операторов, принятые по умолчанию.

4) Возможности ИПЯ: поиск по параметрам документов

Отдельно рассматривается возможность поиска по параметрам документов, которая зачастую позволяет ограничивать диапазон поиска значениями URL, датам, заглавий и т.п. Чаще всего выйти на возможность поиска по параметрам можно из режима расширенного поиска.

В Google, например, обеспечивается поиск по сайту ("site:"), определение ссылок на сайт ("admission site:"), поиск по ценам, например "DVD player $250..350", странам, датам, доменам и т.д. Во многих системах обеспечивается поиск по данным в форматах: HTML, PDF, RTF, MsWord.

5) Адаптивные поисковые интерфейсы

В последнее время получили распространение ***адаптивные интерфейсы уточнения запросов***, чаще всего реализуемые путем кластеризации результатов первичного поиска. Появилось такое понятие, как метод "папок поиска" (Custom Search Folders), который представляет собой множество подходов, общее у которых - попытка сгруппировать результаты поиска и представить кластеры в удобном для пользователей виде.

К подобным механизмам можно отнести, например, австралийский поисковый сервер Mooter (http://www.mooter.com), на котором применяется визуальный подход к предоставлению результатов поиска по обрабатываемым запросам путем группировки результатов первичного поиска по категориям; поисковый сервер iBoogie (http://www.iboogie.com/), где также группирует результаты поиска, но отображает их в виде, близком к экрану проводника Windows. Слова и словосочетания в информационных портретах, применяемых, например, в системах Галактика Zoom и InfoStream, также позволяют адаптивно уточнять первичные запросы. В системе InfoStream информационный портрет расширен многочисленными дополнительными парметрами (персоны, рубрики, страны, размеры документов и т.д).

6) Практические рекомендации

1. Начинайте поиск с наиболее известных и мощных поисковых серверов.
2. Используйте специализированные видовые, тематические или региональные поисковые серверы, если они известны.
3. Внимательно прочтите инструкцию (help, FAQ) по выбранной поисковой системе.
4. Выделяйте ключевые слова для поиска, наиболее точно отражающие интересующую проблематику.
5. Начинайте поиск с простых запросов в режимах простого поиска. По мере получения результатов расширяйте или уточняйте запросы с помощью дополнительных возможностей - используя логические и контекстные операторы, поиск по параметрам, переходя в режим расширенного поиска, используйте возможности адаптивных интерфейсов уточнения первичных запросов.
6. Используйте поиск по параметрам, чаще всего предлагаемый в режимах расширенного поиска - это обеспечит фильтрацию документов по форматам, датам, размерам, странам, языкам и т.д.
7. Если Вас интересует достаточно широкий спектр информации, имеющей отношение к первичному запросу (например, при составлении обзоров), смело используйте режим "поиск подобных документов".
8. В случае наличия средств ранжирования выдачи по релевантности обязательно используйте их. Это позволит Вам достаточно быстро оценить результаты поиска в случаях, когда общие объемы выдачи могут превышать разумные рамки.

Литература

Публикации Д.В. Ландэ по тематике лекции:

1. Григорьев А.Н., Ландэ Д.В. Адаптивный интерфейс уточнения запросов к системе контент-мониторинга InfoStream. // Труды Международного семинара «Диалог'2005» (Звенигород, 1-6 июня 2005 г.). – М.: Наука, 2005. – С. 109-111. (http://poiskbook.kiev.ua/dialog.html).
2. Ландэ Д.В. О чем говорят запросы пользователей к поисковым серверам. // Сети и телекоммуникации. – 1999. - № 4. - С. 19-21.(http://dwl.kiev.ua/art/zap/).

3. Ландэ Д.В. Поиск знаний в Internet. – М.: Диалектика-Вильямс, 2005. (http://poiskbook.kiev.ua).
4. Ландэ Д.В. Искать и не сдаваться. // CHIP/Украина. – 2004. – №5. - С. 84-87 (http://dwl.kiev.ua/art/sart/).

Публикации других авторов:

1. C. Silverstein, M. Henzinger, H. Marais, and M. Moricz. Analysis of a very large altavista query log. Technical Report 1998-014, COMPAQ System Research Center, October 1998.
2. Белоногов Г.Г., Кузнецов Б.А. Языковые средства автоматизированных информационных систем. – М.: Наука, 1983. -288 с.
3. http://help.yandex.ru
4. http://help.rambler.ru

ЛЕКЦИЯ 5. НОВОСТНЫЕ ИНФОРМАЦИОННЫЕ ПОТОКИ В ИНТЕРНЕТ. ИНТЕГРАЦИЯ КОНТЕНТА

1) Новостные потоки в Интернет, СМИ в Интернет, сетевые СМИ

В свое время вместе с переносом в Интернет зародился процесс создания в Сети изданий, электронные версии которых дополняли и заменяли традиционные. СМИ в Интернет зачастую представляют собой прямую репликацию традиционных средств массовой информации на Web-серверах. Однако самые прогрессивные в технологическом плане СМИ смогли "перешагнуть" рамки традиционного представления информации и стать полноценными сетевыми СМИ.

Сетевые СМИ - это новый тип представления информации, изначально ориентированный на Интернет, учитывающий многие нюансы представления информации в этой среде (New Media). Как правило, выпуск традиционными СМИ полноценного сетевого варианта требует не только изменения форматов и формы подачи информации, но и определенной семантической корректировки материалов. Сетевым СМИ присущи такие преимущества, как оперативность, интерактивность, мультимедийность и дешевизна.

В связи с развитием информационных ресурсов сети Интернет документальное информационное пространство развилось до такого уровня, который требует новых подходов. Рост объемов информации и скорости ее распределения фактически породил понятие информационных потоков. Вместе с тем, существующие инструментальные средства уже не всегда способны адекватно отражать ситуацию, речь идет не столько об анализе конечных массивов документов, сколько о навигации в документальных информационных потоках.

2) Проблемы интеграции новостных ресурсов в Интернет

Недостаток HTML: описывает лишь внешний вид Web-сайтов, обеспечивая прежде всего визуализацию данных. Он был разработан исключительно для отображения содержания сайтов, и не всегда удобен для автоматической обработки информации, в том числе и для организации поиска. Т.е. вся сеть Интернет ориентирована на показ пользователям отдельных сайтов и плохо приспособлена для автоматизированного сбора информации, ее классификации и аналитической обработки. Сегодня представление информации на разных сайтах существенно отличаются по оформлению и расположению, что затрудняет ее автоматическую обработку.

Для интеграции контента необходимо использовать унифицированный формат данных на сайтах, стандарт, обеспечивающий однотипный обмен данными в Интернет. В качестве такого унифицированного формата все шире используется язык eXtensible Markup Language (XML) и его диалекты.

3) Интеграторы новостей

Оптимальное решение, способное помочь ориентироваться в новостной информации Интернет, сегодня предоставляют системы синдикации новостей. Под синдикацией в данном

случае понимается сбор информации в Интернет и последующее распространение ее фрагментов в соответствии с потребностями пользователей. Кроме того, службы синдикации обеспечивают публикацию одних и тех же данных на различных сайтах (в том числе, предназначенных для карманных компьютеров и мобильных телефонов).

Технология синдикации Интернет-новостей включает в себя "обучение" программ сбора структуре выбранных источников (Web-сайтов), непосредственное сканирование информации, ее приведение к общему формату (в последнее время - к XML), классификацию и доставку пользователям различными путями (e-mail, Web, WAP, SMS и т.д.).

Примеры служб синдикации новостей:

- MoreOver.com
- NewsIsFree.com
- News.Google.com
- News.Yandex.ru
- Integrum.ru
- WebScan.ru
- UAport.net/UAnews
- InfoStream.ua

4) Форматы синдикации новостей

Для решения задачи синдикации новостей было создано несколько форматов описания данных на основе XML. Самый распространенный формат получил название ***RSS***, что означает Really Simple Syndication, Rich Site Summary, хотя изначально он назывался RDF Site Summary. Смысл всех этих аббревиатур заключается в простом способе обобщения и распределения информационного наполнения Web-сайтов - синдикации контента.

Изначально RSS создавался компанией Netscape для портала Netcenter как одно из первых XML-приложений, но затем стал использоваться на многих других сайтах. Сегодня практически все ведущие новостные сайты. "Живые журналы", работающие в Интернет, используют RSS в качестве инструмента оперативного представления своих обновлений.

Спецификации отдельных версий формата RSS приведены на таких Web-страницах:

- RSS 0.90: http://www.purplepages.ie/RSS/netscape/rss0.90.html
- RSS 0.91: http://my.netscape.com/publish/formats/rss-spec-0.91.html
- RSS 0.92: http://backend.userland.com/rss092
- RSS 0.93: http://backend.userland.com/rss093
- RSS 1.0: http://web.resource.org/rss/1.0/
- RSS 2.0: http://backend.userland.com/rss/

Во всех версиях RSS есть некоторые особенности, но объединяет их ориентация на один тип информации, вследствие чего они содержат общие базовые поля: основной блок данных (channel), который содержит такие атрибуты, как заглавие канала (title), ссылки (link), данные о языке сообщений (language) и логотип (image), после которых идет список самих сообщений, где в каждом пункте (item) указывается заголовок (title), краткое описание (description) и ссылка на новость (link). Кроме того, каждый RSS-файл начинается обязательными элементами xml и rss. Первый из этих элементов содержит атрибуты version (версия) и encoding (кодировка).

Основным применением RSS в настоящее время являются новостные ***фиды*** (feed). Фид - это файл в формате RSS, в который записывается новостной контент Web-ресурса.

Еще один диалект XML - ***OPML*** (Outline Processor Markup Language) используется для описания совокупности RSS-фидов, спецификация которого размещена по адресу http://opml.scripting.com/spec. С помощью OPML обеспечивается эффективный унифицированный обмен списками RSS-фидов.

5) Агрегаторы

Пользователи могут получить доступ к данным в формате RSS с помощью специальных программ, которые в наглядном виде отображают содержание RSS-фидов. Эти программы называются RSS-агрегаторами. Рассматриваются примеры агрегаторов:

- Feedreader (http://www.feedreader.com)
- FeedDemon (www.feeddemon.com)
- Abilon и ActiveRefresh (http://www.activerefresh.com/download.php)
- Syndirella (http://www.yole.ru/projects/syndirella)

Владельцы КПК, установив на свои устройства RSS-агрегаторы, могут эффективно просматривать новостные файлы в RSS - формате. Для платформы Palm OS наиболее популярной является программа компании Stand Alone - Hand RSS (http://standalone.com/palmos/hand_rss/). В качестве еще одного эффективного агрегатора можно назвать программу Quick Palm RSS Reader (http://remus.manilasites.com/). Из специализированных для Pocket PC можно назвать агрегатор новостей в RSS/RDF PocketFeed (http://www.furrygoat.com/Software/).

Литература

Публикации Д.В. Ландэ по тематике лекции:

1. Брайчевский С.М., Ландэ Д.В. Современные информационные потоки: актуальная проблематика. Журнал "Научно-техническая информация", серия 1, № 11, 2005, стр. 21-33 (http://dwl.visti.net/art/nti05/)
2. Ландэ Д.В. Основы интеграции информационных потоков. – К.: Инжиниринг, 2006. – 240 с. (http://dwl.visti.net/art/monogr-osnov/spusk3.pdf)
3. Система интеграции новостей из сети Интернет /Методическое пособие/. Информационный центр «ЭЛВИСТИ», 2003 (http://dwl.visti.net/art/method/broshura.pdf)
4. Ландэ Д.В, Литвин А.Б. Феномены современных информационных потоков // Сети и бизнес. -2001. - № 1. - С. 14-21.(http://dwl.visti.net/art/content/)
5. Д.В. Ландэ, Морозов А.Ю. Редкостный Синтез Сайтов, “Мой Компьютер” №25, 2003 (http://dwl.visti.net/art/rssart/)
6. А.Н. Григорьев, Д.В. Ландэ. New Media – новая информационная среда, "Сети и телекоммуникации", № 4, 2000 (http://dwl.visti.net/art/nm/)
7. Д.В. Ландэ. Об отделении зерен от плевел, “Мой Компьютер” № 33,2003 (http://dwl.visti.net/art/z-p/)
8. Д.В. Ландэ. Информация в фокусе, «Телеком» № 10, 2003 (http://dwl.visti.net/art/inf/index1.html)
9. Д.В. Ландэ, Морозов А.Ю. Новостной Интернет. «Телеком», № 11, 2004, № 1-2, 2005 (http://dwl.visti.net/art/ni/)

Публикации других авторов:

1. Иванов С.А., Круковская Н.В. Статистический анализ документальных информационных потоков // Научно-техническая информация. Информ. процессы и системы. — Сер. 2. — 2004. — № 2. — С. 11–14.
2. Gianna M. Del Corso, Antonio Gullí Univerisity, Francesco Romani. Ranking a stream of news. International World Wide Web Conference. Proceedings of the 14th international conference on World Wide Web. Chiba, Japan. – 2005. - P. 97 - 106.
3. Mark Pilgrim. What is RSS? - 2002. (http://www.xml.com/pub/a/ 2002/12/18/dive-into-xml.html)

ЛЕКЦИЯ 6. МАТЕМАТИЧЕСКИЕ МОДЕЛИ ИНФОРМАЦИОННЫХ ПОТОКОВ

1) Состояние проблемы

Сегодня достигнуты определенные успехи в решении проблемы старения информации в рамках модели Бартона-Кеблера, вместе с тем вопрос динамики информационных потоков остаются почти не исследованными. Динамика сетевых информационных потоков обусловлена многими факторами, большинство из которых не поддаются точному анализу. Однако можно предположить, что общий характер временной зависимости числа тематических публикаций в

Интернет определяется закономерностями, которые целиком допускают построение математических моделей.

2) Баланс тем

Организации-генераторы новостной информации в большинстве работают в стационарном режиме. Это означает, что каждая организацию-генератор производит поток информации, в среднем постоянный по количеству сообщений. Изменяются во времени лишь объемы сообщений, которые соответствуют той или другой теме. Другими словами, рост количества публикаций по одной теме сопровождается уменьшением публикаций на другие темы, так что для каждого промежутка времени T справедливо:

$$\int_0^T \sum_{i=1}^{M} n_i(t)dt = NT ,$$

где $n_i(t)$ – количество публикаций в единицу времени, а M – общее количество всех возможных тем. При моделировании основной интерес представляет изучение динамики отдельного тематического потока, который описывается плотностью $n_i(t)$.

3) Линейная модель

В некоторых случаях динамика тематических информационных потоков (повышение актуальности или старение информации) происходит линейно, т.е. количество сообщений в момент времени t можно представить формулой:

$y(t) = y(t_0) + v(t - t_0)$,

где $y(t)$ – количество сообщений на время t, v – средняя скорость увеличения (уменьшения) интенсивности тематического информационного потока через старение.

Содержательная составляющая информационного потока может быть количественно оцененная как флюктуация информационного потока – изменение стандартного отклонения $\sigma(t)$, которое вычисляется по формуле:

$$\sigma(t_i) = \sqrt{\tfrac{1}{i}\sum_{k=0}^{i}\{y(t_k) - (y(t_0) + v(t_i - t_0))\}^2} .$$

Если эти величины изменяются как корень квадратный из времени, то процесс изменения объемов публикаций по теме можно считать процессом с независимыми приращениями. При этом связями с предыдущими публикациями можно пренебречь.

В случае поведения стандартного отклонения $\sigma(t) \propto t^{\mu}$, чем большее значение μ, тем выше корреляция между текущими и предыдущими сообщениями. В этих случаях μ характеризует степень связи между случайными событиями и принимает значение от ½ до 1.

4) Экспоненциальная модель

В известных работах, посвященных изучению старения информации, используется модель Мальтуса. Преимуществом этой модели заключается в том, что уравнение Мальтуса имеет точное решение в виде очень простой и удобной функции - экспоненты, которая вполне согласуется с практикой для локальных областей.

В этой модели процесс повышения актуальности или старения информации описывается экспонентной зависимостью, которую можно аппроксимировать следующей формулой:

$N(t) = N(t_0)e^{\lambda(t - to)}$,

где λ - среднее относительное изменение интенсивности информационного потока. Вместе с тем, главной проблемой следует считать то, что экспонента является монотонно возрастающей функцией, и поэтому не может описывать процессы, которые по своей природе должны иметь локальные экстремумы.

Относительное изменение интенсивности в определенный момент времени исчисляется по формуле:

$\lambda(t_i)=(N(t_i) – N(t_{i-1}))/N(t_{i-1})$.

Изменение флюктуаций величины $\lambda(t_i)$ относительно среднего значения может быть оценена по формуле:

$$\sigma(t_i) = \sqrt{\frac{1}{i}\sum_{k=0}^{i}\{\lambda(t_k) - \lambda\}^2}.$$

Если $\sigma(t)$ изменяется как корень квадратный из времени, то можно говорить о процессе с независимыми приращениями, корреляции между отдельными сообщениями несущественные. В случае наличия значительной доли зависимых сообщений справедливо: $\sigma(t) \propto t^{\mu}$, причем μ превышает ½, но ограниченно 1.

Значение μ, превышающее ½, говорит о наличии долгосрочной памяти системы. Такие системы порождают класс процессов, получивших название автомодельных, для которых предполагается корреляция между количеством сообщений информационных потоков в разные моменты времени.

Изучение флюктуаций информационных потоков показывает наличие статистической корреляции как на коротких, так и на продолжительных временных интервалах.

5) Логистичексая модель

Логистическую модель можно рассматривать как обобщение экспоненциальной модели Мальтуса, которая, предусматривает пропорциональность скорости роста функции ее значения в каждый момент времени:

$$\frac{dn(t)}{dt} = kn(t),$$

где k – некоторый коэффициент.

В случае логистической модели идея заключается в том, чтобы сделать коэффициент в уравнении Мальтуса функцией времени. Наиболее распространенным есть использования константы, которая в явном виде ограничивает рост решения. В нашем случае с этой целью используем емкость N. Тогда правая часть выражения представляется в виде:

$$k(N - rn(t)),$$

где k – коэффициент Мальтуса, а r – коэффициент, который описывает отрицательные для данной системы процессы, связанные с внутренними факторами.

Вклад интенсивности D определяется следующим образом:

$$y(t) = \begin{cases} D, 0 < t \le \lambda \\ 0, t < 0, t > \lambda \end{cases}$$

Соответственно, рассматриваются две временные области: $0 < t \le \lambda$ с $D > 0$ и $t > \lambda$ с $D = 0$, для которых решениями являются функции $u(t)$ и $v(t)$. Полное решение получается путем “сшивки” на границе в точке λ:

$$n(t) = \begin{cases} u(t), 0 < t \le \lambda \\ v(t), t > \lambda \end{cases}$$

$$u(\lambda) = v(\lambda)$$

Первой области соответствует процесс роста числа публикаций по данной теме в условиях ее ненулевой актуальности ($D > 0$) и, возможно, переход к состоянию насыщения, а второй – процесс сокращения числа публикаций, обусловленный потерей актуальности ($D = 0$).

После нормирования параметров пороговой величины N, представим уравнение для первой области в таком виде:

$$\frac{du(t-\tau)}{dt} = pu(t-\tau)(1 - qu(t-\tau)) + Du(t-\tau),$$

$$u(0) = n_0$$

Решение этого уравнения имеет вид:

$$u(t) = \frac{u_s}{1 + (\frac{u_s}{n_0} - 1)\exp[-(p + D)(t - \tau)]}$$

Для второй области, соответственно, имеем:

$$\frac{dv(t-\lambda)}{dt} = pv(t-\lambda)(1 - qv(t-\lambda)),$$

$$v(\lambda) = u(\lambda)$$

Для нее решение имеет такой вид:

$$v(t) = \frac{u(\lambda)}{qu(\lambda) + (1 - qu(\lambda))\exp[-p(t-\lambda)]}.$$

Данная модель на определенных участках t совпадает с линейной и экспоненциальной моделями.

6) Перспективные направления в моделировании

Представленные модели соответствуют большинству из наблюдаемых на практике распределений объемов тематических публикаций. Вместе с тем исследовательские данные свидетельствуют о наличии еще нескольких типов зависимостей.

Отдельную проблему информационной динамики представляют циклические процессы роста и снижения активности информационных потоков, не связанные с информационными факторами (например, периодическое снижение количества публикаций по выходным дням).

Литература

Публикации Д.В. Ландэ по тематике лекции:

1. Ландэ Д.В. Некоторые методы анализа новостных информационных потоков. // Научные труды Донецкого национального технического университета. Серия: Информатика, кибернетика и вычислительная техника (ИКВТ-2005). - Вып. 93. – Донецк: ДонНТУ, 2005. - С. 277-287. (http://dwl.visti.net/art/don/) .
2. Ландэ Д.В. Основы интеграции информационных потоков. – К.: Инжиниринг, 2006. – 240 с. (http://dwl.visti.net/art/monogr-osnov/spusk3.pdf)
3. Брайчевский С.М., Ландэ Д.В. Современные информационные потоки: актуальная проблематика // Научно-техническая информация. Сер. 1. –М., 2005. - №11, - С. 21-33. (http://dwl.visti.net/art/nti05/).

Публикации других авторов:

1. В.И. Арнольд. Аналитика и прогнозирование: математический аспект. // Научно-техническая информация. Сер. 1. Вып. 3. - 2003. - С. 1-10.
2. Иванов С.А., Круковская Н.В. Статистический анализ документальных информационных потоков. // Научно-техническая информация. Сер. 2. Вып. 2. – 2004. - С. 11-14.
3. Gianna M. Del Corso, Antonio Gullí Univerisity, Francesco Romani. Ranking a stream of news. International World Wide Web Conference. Proceedings of the 14th international conference on World Wide Web. Chiba, Japan. – 2005. - P. 97 - 106.
4. Ефимов А.Н. Информация: ценность, старение, рассеяние. М., 1978.
5. Мотылев В. М. Старение научно-технической литературы. – Л., Наука, 1986.
6. Горькова В. И. Информетрия (Количественные методы в научно-технической информации) //Итоги науки и техники. Сер. Информатика. Т.10. – М.: ВИНИТИ, 1988.
7. Вольтерра В. Математическая теория борьбы за существование. М.: Наука, 1976.

ЛЕКЦИЯ 7. КЛАСТЕРНЫЙ АНАЛИЗ И ИНФОРМАЦИОННЫЙ ПОИСК

Кластерный анализ - метод группировки экспериментальных данных в классы. Наблюдения, попавшие в один класс, в некотором смысле ближе друг к другу, чем к наблюдениям из других классов. *(Глоссарий.ru)*

1) Понятие «информационного портрета»

Вводится понятие информационного портрета темы как набора термов, а также некоторых весовых коэффициентов этих термов. Показывается, как при определенных заранее информационных портретах тематических рубрик, для любого документа могут быть вычислены его веса в пространстве тематических рубрик.

Приводится алгоритм взвешивания потока документов в пространстве информационных портретов.

Преимущество подхода: два в одном

Чаще всего информационные портреты формируются путем лингвостатистического анализа массивов документов, полученных в результате поиска по соответствующим тематическим запросам. Эти запросы в большинстве промышленных информационно-поисковых систем составляются на языках, являющихся расширением булевой алгебры.

Окончательная же рубрикация документов предполагает более «экономный» весовой подход на основе массива термов, полученных в результате периодической отработки булевых запросов. Таким образом, в результате учитываются «логические» преимущества первого подхода и эксплуатационные – второго.

2) Взаимосвязь термов в информационном портрете

Пусть операция отображения потока документов в пространство информационных портретов, задается матрицей M. Введем понятие ядра этой операции как произведения матриц $A = M^T M$. Матрица A по смыслу представляет собой матрицу взаимосвязей информационных портретов.

Еще одна матрица, полученная в результате умножения $B = M M^T$ выражает взаимосвязь документов. Для современных информационных потоков размерность матриц этого типа намного превышает размерность матриц A. Выявляя явные группы взаимосвязанных тем в матрице A, можно определять группы взаимосвязанных документов в матрице B, блочная группировка которой ввиду ее размерности и динамики роста весьма затратная.

3) Латентное семантическое индексирование

Метод кластерного анализа LSI (латентного семантического индексирования), базируется на сингулярном разложении матриц (SVD). Сингулярным разложением матрицы A называется ее разложение вида $A=USV^T$, где U и V – ортогональные матрицы, а S – диагональная матрица, элементы которой $s_{ij} = 0$, если i не равно j, а $s_{ii} >= 0$. Величины s_{ii} называются сингулярными числами матрицы A.

В рассматриваемом примере (таблиц взаимосвязей) матрица $A = M^T M$ – квадратная, однако метод LSI применяется и к прямоугольным матрицам, но в этих случаях размерность матрицы S соответствует рангу матрицы A.

В соответствии с методом LSI в рассмотрение берутся k наибольших сингулярных значений, а каждому такому сингулярному значению матрицы A соответствует кластер взаимосвязанных тем. Таким образом матрица A в соответствии с нормой Фрбениуса аппроксимируется матрицей $A_k = \Sigma u_i s_{ii} v_i^T$.

Метод LSI применим и к ранжированию выдачи информационно-поисковых систем, основанному на цитировании. Это алгоритм HITS (Hyperlink Induced Topic Search) – один из двух самых популярных на сегодня в области информационного поиска.

Ввиду своей вычислительной трудоемкости (равной $O(N^2)$, N – размерность A), этот метод LSI применяется только для относительно небольших матриц.

4) Взаимосвязь рубрик и метод k-means

Рассмотрим множество уникальных термов из всех тематических информационных портретов W и проекцию (P) множества информационных портретов на W.

Тогда произведение матриц $E = P^T P$ будет таблицей взаимосвязей тем, построенной в результате анализа состава термов соответствующих информационных портретов. Укрупнение рубрик – актуальная задача кластерного анализа и она может быть решена путем их группировки по признакам подобия.

Показывается, как можно выделить некоторое число групп взаимосвязанных рубрик методом кластерного анализа k-means.

Существуют две реализации алгоритма k-means, «жесткая», когда число k фиксировано и «мягкая», которая позволяет на основании некоторых критериев оценить значение k. Суть жесткого алгоритма k-means определяется следующим образом: случайным образом выбирается k векторов-строк, которые определяются как центроиды (наиболее типичные представители) кластеров. Затем k кластеров наполняются – для каждого из оставшихся векторов-строк определяется близость к центроиду соответствующего кластера. После этого вектор-строка приписывается к тому кластеру, к которому он наиболее близок. После этого строки-векторы группируются и перенумеровываются в соответствии с полученной группировки. Затем для каждого из новых кластеров заново определяется центроид – вектор-строка, наиболее близкая ко всем векторам из данного кластера (например, тот, сумма скалярных произведений которого с каждым из векторов кластера - минимальна). После этого заново выполняется процесс наполнения кластеров, затем вычисление новых центроидов и т.д., пока процесс формирования кластеров не стабилизируется (или набор центроидов не повторится).

В отличие от метода LSI, k-means идеально подходит для кластеризации динамических информационных потоков.

5) Гибридный метод - выявление сюжетов

Если к потоку документов за некоторый промежуток время добавляется несколько новых документов, то каждый из новых документов приписывается к соответствующему кластеру в соответствии с некоторой мерой близости, после чего происходит пересчет центроидов. Рекурсивный пересчет наполнения кластеров в соответствии с новыми центроидами может выполняться с заданной заранее периодичностью.

Алгоритм выявления основных сюжетных цепочек, используемый в системе InfoStream заключается в следующем. Последний поступивший на вход системы документ (документ с номером 1 при обратной нумерации) порождает первый кластер и сравнивается со всеми предыдущими в соответствии с некоторой метрикой μ. Если эта мера близости для какого-нибудь документа оказывается ближе заданной пороговой, то текущий документ приписывается первому кластеру. Сравнение продолжается, пока не исчерпывается список актуальных документов потока. После такой обработки документа 1, происходит обработка следующего документа, не вошедшего в первый кластер, с которым последовательно сравниваются все актуальные документы потока и т.д. В результате формируется некоторое неизвестное заранее количество кластеров, которые ранжируются по своим весам. Для выбранных кластеров заново пересчитываются центроиды – документы наиболее отражающие тематику кластера.

6) Применение методов кластерного анализа при построении систем интерактивного первично введенных запросов

Названный выше метод "папок поиска" представляет собой множество подходов с использованием кластерного анализа, общее у которых - попытка сгруппировать результаты поиска и представить кластеры в удобном для пользователей виде для дальнейшей навигации и уточнения запросов.

Литература

Публикации Д.В. Ландэ по тематике лекции:

1. Ландэ Д.В. Некоторые методы анализа новостных информационных потоков. // Научные труды Донецкого национального технического университета. Серия: Информатика, кибернетика и вычислительная техника (ИКВТ-2005). - Вып. 93. – Донецк: ДонНТУ, 2005. - С. 277-287. (http://dwl.visti.net/art/don/).
2. Ландэ Д.В. Основы интеграции информационных потоков. – К.: Инжиниринг, 2006. – 240 с. (http://dwl.visti.net/art/monogr-osnov/spusk3.pdf)
3. Ландэ Д.В. Поиск знаний в Internet. –М.: Диалектика-Вильямс, 2005. (http://poiskbook.kiev.ua)

Публикации других авторов:

1. А.В. Антонов. Методы классификации и технология Галактика-Зум. // Научно-техническая информация. Сер. 1. Вып. 6. - 2004. - С. 20-27.
2. Кириченко К.М, Герасимов М.Б. Обзор методов кластеризации текстовых документов // Материалы международной конференции Диалог'2001, (http://www.dialog-21.ru/Archive/2001/volume2/2_26.htm)
3. Мандель И.Д. Кластерный анализ. - М.: Финансы и статистика, 1988. - 176 с.
4. Сегалович И.В. Как работают поисковые системы. // Мир Internet. – 2002. -№ 10. (http://www.dialog-21.ru/direction_fulltext.asp?dir_id=15539).
5. Некрестьянов И.С., Добрынин В.Ю., Клюев В.В. Оценка тематического подобия текстовых документов // Труды второй всероссийской научной конференции “Электронные библиотеки”. – Протвино, 2000. – С. 204-210.
6. Landauer, T.K., Foltz, P.W., Laham, D. An introduction to latent semantic analysis. - Discourse Processes, - V 25, - 1998. P.. 259-284.
7. Chakrabarti Soumen. Mining the web. Discovery knowledge from hypertext data. - Publisher: Morgan Kaufmann, 2002. - 344 p.
8. G.W. Furnas, S. Deerwester, S.T. Dumais, T.K. Landauer, R. A. Harshman, L.A. Streeter, and K.E. Lochbaum. Information retrieval using a Singular Value Decomposition Model of Latent Semantic Structure. - ACM SIGIR, 1988.
9. Thomas Hofmann. Probabilistic latent semantic indexing. In Proc. of the SIGIR'99. – 1999. - P. 50-57

ЛЕКЦИЯ 8. РАНЖИРОВАНИЕ РЕЗУЛЬТАТОВ ПОИСКА

Ранжирование - процесс, при котором поисковая система:
- принимает запрос пользователя;
- находит все подходящие веб-страницы; и
- выстраивает их в определенном порядке по принципу наибольшего соответствия конкретному запросу.

Выведение рейтинга зависит от алгоритма ранжирования, которым пользуется поисковая машина. *(Глоссарий.ru)*

1) Задача ранжирования

В результате поиска пользователь может получить обширный список релевантных документов. Сортировка этого списка таким образом, чтобы наиболее важные для пользователя документы были в начале этого списка, в технологиях ИПС принято называть ранжированием результатов поиска или ранжированием откликов ИПС.

Необозримость возможного списка приводит к идее сортировки всей базы данных по определенным параметрам, которая заменяет поиск.

Сортировка результатов поиска по уровню релевантности подходит не для всех моделей поиска (например, не подходит для булевой модели).

Перспективный путь – использование многопрофильных шкал, сформированных на основе метаданных, использование кластерного анализа.

Реализация сюжетных цепочек в тематических информационных массивах и их взвешивание рассматриваются как один из алгоритмов ранжирования.

2) Особенности ранжирования текстовых и гипертекстовых документов

- Ранжирование гипертекстовых документов по ссылкам.
- Ранжирование текстовых документов - по уровню релевантности и другим параметрам (времени публикации, авторитетности источника, автора).

3) Ранжирование с учетом гиперсвязей - HITS и PageRank

HITS (Hyperlink Induced Topic Search):

Применение метода латентного семантического индексирования к ранжированию выдачи информационно-поисковых систем, основанному на цитировании.

Алгоритм HITS обеспечивает выбор из информационного потока лучших «авторов» (первоисточников) и «посредников» (документов от которых идут ссылки цитирования). Понятно, что страница является хорошим посредником, если она содержит ссылки на ценные первоисточники, и наоборот, страница является хорошим первоисточником, если она упоминается хорошими посредниками.

Для каждого документа рекурсивно вычисляется его значимость как первоисточника $\boldsymbol{a_p}$ и посредника $\boldsymbol{h_p}$ по формулам:

$$\boldsymbol{a_p = \Sigma h_q,\ h_p = \Sigma a_q.}$$

Если ввести понятие матрицы инциденций A, элемент которой a_{ij} равен единице, если документ D_i содержит ссылку на документ D_j, и нулю в противном случае, то алгоритм HITS обеспечивает выбор наиболее авторитетных документов документов (первоисточников – $\boldsymbol{a_p}$ или посредников - $\boldsymbol{h_p}$), которые предположительно соответствуют собственным векторам матриц AA^T и A^TA с наибольшими модулями собственных значений. В этом смысле алгоритм HITS эквивалентен LSI.

Действительно, пусть, в соответствии с сингулярным разложением $A = USV^T$, S – квадратная диагональная матрица. Тогда $AA^T = USV^T\ VSU^T = US\ I\ SU^T = US^2U^T$, где S^2 – диагональная матрица с элементами s_{ii}^2. Очевидно, как и при LSI, собственные векторы, соответствующие наибольшим сингулярным значениям AA^T и/или A^TA будут соответствовать наиболее статистически важным авторам и/или посредникам.

Алгоритм вычисления рангов HITS влечет рост рангов страниц при увеличении количества и степени связанности страниц соответствующего сообщества. В этом случае в результат выдачи информационно-поисковой системы может попасть много страниц на темы, не соответствующие информационной потребности пользователя, то есть часть выдаваемых результатов, соответствующих требуемой теме, может оказаться не доминирующей. Это обуславливает присвоение высших рангов страницам на тему, не требуемую пользователем, т.е. происходит смещение тематики (topic drift).

Как некоторое расширение стандартного алгоритма HITS рассматривается алгоритм PHITS - Probabilistic HITS, авторы - Д. Кон (D. Cohn) и Х. Чанг (H. Chang). Предполагается: D – множество цитирующих документов, C – множество ссылок, Z- множество классов (факторов). Пусть $d \in D$, генерируется с вероятностью $P(d)$.

Условные вероятности $P(c|z_k)$ и $P(z_k|d)$ для описания зависимостей между наличием ссылки $c \in C$, латентным фактором $z_k \in Z$ и документом - d.

Оценивается (максимизируется функция):

$$L(D,C) = \prod P(d,c) = \prod_{c \in C,\ d \in D} P(d)P(c|d),$$

где

$$P(c|d) = \Sigma_k\ P(c|z_k)\ P(z_k|d).$$

Задача состоит в том, чтобы подобрать $P(z_k)$, $P(c|z_k)$, $P(z_k|d)$, чтобы максимизировать $L(D,C)$.

После этого:

$P(c|z_k)$ – ранги первоисточников

$P(d|z_k)$ – ранги посредников.

Для вычисления рангов необходимо задать количество факторов в Z, и тогда $P(c|z_k)$ будет характеризовать качество страницы как “первоисточника” в контексте тематики z_k. Кроме того, итеративный процесс зачастую останавливается не на абсолютном, а на локальном максимуме функции правдоподобия L.

В ситуациях, когда в множестве Web-страниц нет явного доминирования тематики запроса PHITS ведет себя лучше HITS.

PageRank:

Алгоритм PageRank близок по идеологии к литературному индексу цитирования и рассчитывается с учетом количества ссылок из других документов на данный документ. PageRank, в отличие от литературного индекса цитирования, не считает все ссылки равными.

Объяснение принципа подсчета ранга Web-страницы PageRank следующее. Рассматривается процесс, при котором пользователь сети Internet открывает случайную Web-страницу, из которой переходит по случайно избранной гиперссылке. Потом он перемещается на другую Web-страницу и снова активизирует случайную гиперссылку и т.д., постоянно переходя от странице к странице, никогда не возвращаясь. Иногда ему такое блуждание надоедает, и он снова переходит на случайную Web-страницу - не по ссылке, а набрав вручную некоторый URL. В этом случае, вероятность того, что блуждающий в Сети пользователь перейдет на некоторую определенную Web-страницу - это ее ранг - PageRank. PageRank Web-страницы тем выше, чем больше других страниц ссылается на нее, и чем эти страницы популярнее.

Пусть есть n страниц $T=\{T_1, T_2, ..., T_n\}$, которые ссылаются на данный документ (Web-страницу A) , а $C(A)$ — общее число ссылок с Web-страницы A на другие документы. Пусть d (damping factor) - это вероятность того, что пользователь, пересматривая какую-нибудь Web-страницу из множества T, перейдет на страницу A по ссылке, а не набирая ее URL или другим средствами. Тогда вероятность продолжения Web-серфинга не используя гиперссылок, путем ручного введения адреса (URL) из случайной страницы будет составлять $1-d$. Индекс PageRank $PR(A)$ для страницы A вычисляется по формуле:

$$PR(A) = (1-d) + d\,(PR(T1)/C(T1) + ... + PR(Tn)/C(Tn)).$$

Таким образом индекс легко подсчитывается простым итерационным алгоритмом.

Несмотря на расхождение данных алгоритмов, в них общее то, что авторитетность (вес) узла зависит от веса других узлов, а уровень "посредника" зависит от того, насколько авторитетные соседние узлы. Кроме того, оба алгоритмы используют вычисление собственных векторов для матриц взаимосвязи (инциденций) соответствующих Web-страниц. Расчет авторитетности отдельных документов сегодня широко используется в таких применениях, как определение порядка сканирования документов, ранжирование результатов поиска, формирование тематических сюжетов и т.п. Формулы расчета авторитетности постоянно совершенствуются. Предполагается, что применение этих алгоритмов в будущем станет еще более эффективным, так как гиперссылки между документами постоянно оптимизируются с учетом предпочтений пользователей и ориентируясь на существующие методы их обработки поисковыми системами.

4) Модель выявления основных и маргинальных тем в новостных потоках

В этом алгоритме ключевые слова (устойчивые словосочетания) из сообщений или отдельные сообщения информационных потоков, порождаемых информационными Web-сайтами выступают аналогами дискретных сигналов. Каждому сообщению приписывается вес, который равен усредненной частоте появления во всем информационном потоке значимых ключевых слов. Очевидно, чем меньше этот вес, тем документ более уникальный.

Рассматривается двухпроходный алгоритм формирования словаря уникальных слов из входного массива из N сообщений (первый проход), а также весов отдельных сообщений (второй проход). Вес сообщения определяется по формуле:

$$W_D = \frac{\sum_{w \in D} w}{|D|}.$$

где W_D – вес сообщения, w – ключевое слово из сообщение, $|D|$ - количество ключевых слов в документе. В рамках модели как вес ключевых слов употребится частота их появлений во входном информационном потоке. В свою очередь, эта частота зависит от объема самого потока и от количества уникальных слов, т.е. объема автоматически формированного словаря уникальных слов.

5) Ранжирование новостных сайтов «по Хиршу»

В 2005 г. в области наукометрии произошло важное событие - физиком Йоргом Хиршем был предложен новый метод оценки научных публикаций, претендующий на более высокую точность и, что особенно важно, объективность по сравнению с получившим широкое распространение индексом цитирования.

Метод состоит в подсчете числа h публикаций одного автора, на которые имеется не менее h ссылок. Учёный имеет индекс h, если h из его N_p статей цитируются как минимум h раз каждая, в то время как оставшиеся $(N_p - h)$ статей цитируются менее чем h раз каждая. На практике общая цитируемость соотносится с цитируемостью по Хиршу следующим образом:

$N_{tot} \sim a\, h^2$.

Параметр Хирша для сайта-источника равен максимальному количеству дней в месяце (h), в течение которых было зафиксировано не менее h внешних ссылок на данный сайт (параметр на практике подсчитывается для новостных сайтов).

Рейтинг Хирша характеризует как регулярность ссылок на источники, так и количество этих ссылок. Показатель Хирша учитывает стабильность авторитетности источника на протяжении длительного периода.

Литература

Публикации Д.В. Ландэ по тематике лекции:

1. Ландэ Д.В. Подход к анализу новостных потоков как дискретных сигналов. // Регистрация, хранение и обраб. данных. – К., 2006. - Т. 8, № 1. - С. 67 - 73. (http://dwl.visti.net/art/ts/).
2. Ландэ Д.В. Основы интеграции информационных потоков. – К.: Инжиниринг, 2006. – 240 с. (http://dwl.visti.net/art/monogr-osnov/spusk3.pdf)
3. Поиск знаний в Internet. Ландэ Д.В. – М.: Диалектика-Вильямс, 2005. (http://poiskbook.kiev.ua).
4. Снарский А.А., Ландэ Д.В., Григорьев А.Н., Брайчевский С.М., Дармохвал А.Т. Ранжирование сайтов «по Хиршу». // Доклады международной конференции «MegaLing'2006 Горизонты прикладной лингвистики и лингвистических технологий». 20-27 сентября 2006, Украина, Крым, Партенит. - с. 248-249.

Публикации других авторов:

1. Сегалович И.В. Как работают поисковые системы. // Мир Internet. – 2002. -№ 10. (http://www.dialog-1.ru/direction_fulltext.asp?dir_id=15539).
2. S. Brin and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. WWW7, - 1998.
3. J.M. Kleinberg. Authoritative sources in a hyperlink environment. // In Processing of ACM-SIAM Symposium on Discrete Algorithms, 1998, 46(5):604–632.
4. Gianna M. Del Corso, Antonio Gullí Univerisity, Francesco Romani. Ranking a stream of news. International World Wide Web Conference. Proceedings of the 14th international conference on World Wide Web. Chiba, Japan. – 2005. - P. 97 - 106.

ЛЕКЦИЯ 9. ЭЛЕМЕНТЫ ФРАКТАЛЬНОГО АНАЛИЗА ИНФОРМАЦИОННЫХ ПОТОКОВ

1) Понятие «фрактал»

Термин *фрактал*, был предложен Б. Мандельбротом в 1975 году для обозначения нерегулярных самоподобных математических структур. Основное определение фрактала, данное Мандельбротом, звучало так: *"Фракталом называется структура, состоящая из частей, которые в каком-то смысле подобны целому"*.

Главная особенность фракталов заключается в том, что их размерность не укладывается в привычные геометрические представления. Фракталам характерна геометрическая «изрезанность». Поэтому используется специальное понятие фрактальной размерности, введенное Ф. Хаусдорфом и А. Безиковичем. Размерность фракталов – не является целым числом, характерным для привычных геометрических объектов

2) Примеры абстрактных фракталов

Алгоритм построения множества Мандельброта основан на итеративном вычислении по формуле:

$$\boldsymbol{Z}[i+1] = \boldsymbol{Z}[i] * \boldsymbol{Z}[i] + \boldsymbol{C},$$

где $\boldsymbol{Z}$ и $\boldsymbol{C}$ - комплексные переменные.

Приводится алгоритм построения и другого фрактального множества - снежинки Коха.

3) Фракталы в природе

Один из лучших примеров проявления фракталов в природе –структура береговых линий. Действительно, на километровом отрезке побережье выглядит столь же изрезанным, как и на стокилометровом. Опыт показывает, что длина береговой линии L зависит от масштаба l, которым проводятся измерения, и увеличивается с уменьшением последнего по степенному закону $L = \Lambda\, l^{1-\alpha}$, $\Lambda = const$. Так, например, для побережья Великобритании $\alpha \approx 1.24$, т.е. фрактальная размерность береговой линии Великобритании равна 1.24 .

4) Информационное пространство и фракталы

В настоящее время информационное пространство принято рассматривать как стохастическое. Во многих моделях информационного пространства изучаются структурные связи между тематическими множествами, входящими в это пространство. Самоподобие информационного пространства выражается, прежде всего в том, что при его лавинообразном росте, частотные и ранговые распределения, получаемые в таких разрезах, как источники, авторы, тематика практически не меняют своей формы. Применение теории фракталов при анализе информационного пространства позволяет с общей позиции взглянуть на закономерности, составляющие основы информатики. Например, тематические информационные массивы сегодня представляют развивающиеся самоподобные структуры, которые по своей сути являются стохастическими фракталами, так как их самоподобие справедливо на лишь уровне математических ожиданий, как например, распределения кластеров по размерам.

В информационном пространстве возникают, формируются, растут и размножаются кластеры – группы взаимосвязанных документов. Системы, основанные на кластерном анализе, самостоятельно выявляют новые признаки объектов и распределяют объекты по новым группам.

Фрактальные свойства характерны для кластеров информационных Web-сайтов, на которых публикуются документы, соответствующие определенным тематикам. Эти кластеры, как наборы тематических документов, представляют собой фрактальные структуры, обладающие рядом уникальных свойств. Например, определена фрактальная размерность подобных информационных массивов, изменяющаяся в пределах от 1.05 до 1.50, что свидетельствует о небольшой плотности заполнения кластеров документами по одной теме.

Свойства самоподобия фрагментов информационного пространства наглядно демонстрирует новый интерфейс представленный на веб-сайте службы News Is Free (http://newsisfree.com). На этом сайте отображается состояние информационного пространства в виде ссылок на источники и отдельные сообщения. В рамках этой модели можно наблюдать «дробление» групп источников при увеличении ранга популярности и «свежести» изданий.

Топология и характеристики моделей Web-пространства оказалываются примерно одинаковыми для различных подмножеств, подтверждая тем самым наблюдение о том, что "Web - это фрактал", т.е. свойства структуры всего Web-пространства Bow Tie также верны и его отдельных подмножеств.

5) Фракталы и временные ряды

Объемы сообщений в тематических информационных потоках образуют временные ряды. Для исследования временных рядов сегодня все шире используется теория фракталов. Временные ряды, порождаемые тематическими информационными потоками, обладают фрактальными свойствами и могут рассматриваться как стохастические фракталы. Этот подход расширяет область применения теории фракталов на информационные потоки, динамика которых описывается средствами теории случайных процессов.

Фрактальная размерность в кластерной системе, соответствующей тематическим информационным потокам, показывает степень заполнения информационного пространства сообщений в течение определенного времени.

6) Показатель Херста

Сегодня в связи с развитием теории стохастических фракталов становится популярной такая характеристика временных рядов как показатель Херста (*H*). Показатель Херста связывают с коэффициентом нормированного размаха (*R/S*), где *R* — вычисляемый определенным образом «размах» соответствующего временного ряда, а *S* — стандартное отклонение.

В свое время Херст экспериментально обнаружил, что для многих временных рядов справедливо: $R / S = (N / 2)^H$. Показатель Херста связан с традиционной «клеточной» фрактальной размерностью (*D*) простым соотношением:

$$D = 2 - H.$$

Известно, что показатель Херста представляет собой меру персистентности — склонности процесса к трендам (в отличие от обычного броуновского движения). Значение $H > ½$ означает, что направленная в определенную сторону динамика процесса в прошлом, вероятнее всего, повлечет продолжение движения в том же направлении. Если $H < ½$, то прогнозируется, что процесс изменит направленность. $H = ½$ означает неопределенность — броуновское движение.

7) Описание вычислительного эксперимента

В качестве экспериментальной базы для исследования фрактальных свойств тематических информационных потоков использовалась система контент-мониторинга InfoStream. Тематика исследуемого информационного потока определялась запросом к системе InfoStream. Исходные данные были получены из интерфейса режима «Динамика появления понятий». На основании обработки этих данных была получена полная картина экспериментальных данных — временной ряд за указанный период.

Для значения $N = 424$ дней по формуле был вычислен показатель Херста, который оказался равным 0,62, что свидетельствует о положительной персистентности всего временного ряда, соответствующего запросу.

Изучение такой характеристики как показатель Херста позволяет прогнозировать динамику информационных потоков, сообщения которых отражают процессы, происходящие в реальном мире.

Приведенные в примере данные подтвердили лежащее в основе исследования предположение об итеративности процессов в информационном пространстве, наличие статистической корреляции в информационных потоках на длительных временных интервалах. Републикации,

цитирование, прямые ссылки и т.п. порождают самоподобие, проявляющееся в устойчивых статистических распределениях и известных эмпирических законах.

Литература

Публикации Д.В. Ландэ по тематике лекции:

1. Ландэ Д.В. Фракталы и кластеры в информационном пространстве. // Корпоративные системы. –2005. - №6, 2005. - С. 35-39. (http://dwl.visti.net/art/frak/)
2. Ландэ Д.В. Основы интеграции информационных потоков. – К.: Инжиниринг, 2006. – 240 с. (http://dwl.visti.net/art/monogr-osnov/spusk3.pdf)
3. Поиск знаний в Internet. Ландэ Д.В. – М.: Диалектика-Вильямс, 2005. (http://poiskbook.kiev.ua)
4. Ландэ Д.В. Фрактальные свойства тематических информационных потоков из Интернет. // Регистрация, сбор и обработка данных, - К., 2006. – Т 8. - № 2. – С. 93 – 99 (http://dwl.visti.net/art/frak-ip/)

Публикации других авторов:

1. Б. Мандельброт. Фрактальная геометрия природы. – М.: Институт компьютерных исследований, 2002 г. - 656 с.
2. Б. Мандельброт. Фракталы, случай и финансы. – М.: Регулярная и хаотическая динамика, 2004 г. - 256 с.
3. *Федер Е.* Фракталы. —М.: Мир, 1991. — 254 с.
4. Э. Петерс. Хаос и порядок на рынках капитала. Новый аналитический взгляд на циклы, цены и изменчивость рынка: Пер. с англ. - М.: Мир. 2000. -333 с.
5. *Иванов С.А.* Стохастические фракталы в Информатике // Научно-техническая информация. — Сер. 2. — 2002. — № 8. — С. 7–18.
6. *Van Raan A.F.J.* Fractal Geometry of Information Space as Represented by Cocitation Clustering // Scientometrics. —1991. — Vol. 20, N 3. — P. 439–449.

ЛЕКЦИЯ 10. ОСНОВНЫЕ СВЕДЕНИЯ О НЕЙРОННЫХ СЕТЯХ

1) Основные сведения

Упрощенная модель мозга человека, содержащая 10^{11} нейронов.

Состав нейрона: аксон, 10 000 дендритов, синапсы.

Нейрон как компьютер: потенциал нейрона (аксона) – функция от потенциала дендритов.

Состояния нейрона – возбужденное или невозбужденное, зависит от величины потенциала.

Формальный нейрон. Входные сигналы формируются в рецепторах (не входят в нейрон). Далее эти сигналы умножаются на веса соответствующих синапсов (которые могут изменяться при обучении), затем результаты суммируются. На основе полученной суммы, полученной с помощью активационной функции, вычисляется выходной сигнал нейрона.

Примеры активационных функций:

1. OUT = K*NET
2. OUT= 1, NET > T
   OUT = 0, NET≤T
3. OUT = 1 / (1 + e^{-NET})
4. OUT = th(NET) …

2) Нейронная сеть, перцептрон

Нейронная сеть – ориентированный ациклический граф, вершины которого нейроны разбиты на слои. Ребра – синапсы. Каждому ребру приписан свой вес и функция проводимости.

Нейронная сеть, способная обучаться.

Первая версия перцептрона представляла собой однослойную нейронную сеть. Однослойный перцептрон состоит из слоя рецепторов и слоя нейронов (внутри которых выделяют слой синапсов). В перцептроне каждый нейрон связан через синаптический контакт со всеми рецепторами предыдущего слоя. Перцептрон представляет интерес с точки зрения его адаптивности в задачах распознавания образов. Было доказано, что однослойные нейронные сети не способны решать многие задачи (например, исключающее ИЛИ). Для решения этих проблем широко используются многослойные нейронные сети. Многослойные сети могут образовываться каскадами слоев. Выход одного слоя является входом для последующего слоя.

Нейронная сеть обучается, чтобы для некоторого множества входных сигналов давать желаемое множество выходных сигналов. Каждое множество сигналов при этом рассматривается как вектор. Обучение осуществляется путем последовательного предъявления входных векторов с одновременной подстройкой весов в соответствии с определенной процедурой. В процессе обучения веса сети постепенно становятся такими, чтобы каждый входной вектор вырабатывал требуемый выходной вектор, используя правила, указанные выше.

3) Пример практического применения: определение тональности сообщений

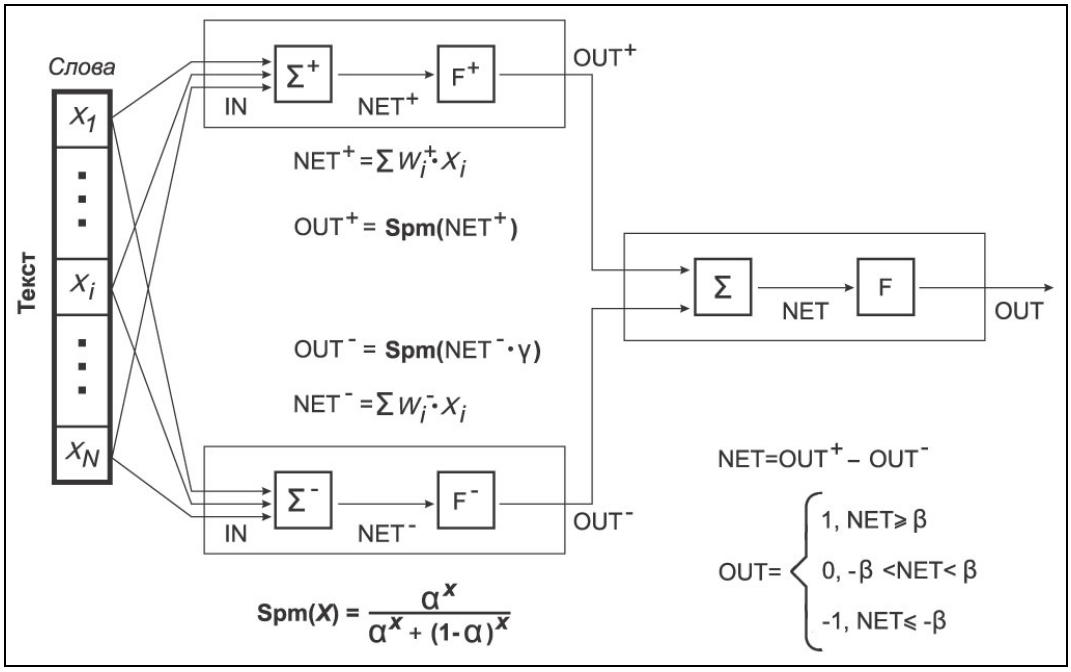
Алгоритм определения тональности представим в виде нейронной сети. Первый слой этой сети составляют два нейрона – определители весовых значений положительной и отрицательной тональности (положительный и отрицательный нейроны). Можно предположить, что количество дендритов каждого нейрона равно количеству слов из словаря естественного языка. На вход нейронов поступают входные сигналы - значения $x_1 \ldots x_n$, соответствующие входным словам. При этом $x_i=1$, если на вход поступило слово из словаря с номером i, в противном случае $x_i=0$. Весовые значения (веса синапсов), которые соответствуют этим словам, равны $w^+_1 \ldots w^+_n$ для положительного нейрона и $w^-_1 \ldots w^-_n$ - для отрицательного. Именно эти весовые значения могут изменяться в процессе обучения перцептрона. Сумматоры подсчитывают значения NET^+ и NET^-, соответственно. Проводимость нейронов рассчитывается по формуле:

$$Spm(x) = \frac{\alpha^x}{\alpha^x + \lambda(1-\alpha)^x}$$

где x – число весомых с точки зрения тональности слов в информационном сообщении, α – вес. Аргументом в этой формуле выступает значение NET^+ для положительного нейрона и γ NET^- – для отрицательного. Оба нейрона выдают через аксоны градиентные значения, OUT^+ и OUT^-, которые являются входными сигналами для нейрона второго уровня, сумматор которого вычисляет разность OUT^+ и OUT^-, а функция проводимости выдает градиентный результат по условию, приведенному на рисунке:

Литература

1. Уссермен Ф. Нейрокомпьютерная техника. - М.: Мир, - 1992. – 184 с.
2. Ландэ Д.В. Основы интеграции информационных потоков. – К.: Инжиниринг, 2006. – 240 с. (http://dwl.visti.net/art/monogr-osnov/spusk3.pdf)
3. Рычагов М. Н. Нейронные сети: многослойный перцептрон и сети Хопфилда. // EXPonenta Pro. Математика в приложениях, N 1, 2003 (http://nature.web.ru/db/msg.html?mid=1193685).
4. Роберт Хехт-Нильсен. Нейрокомпьютинг: история, состояние, перспективы. // Открытые системы. - № 4. -1998. (http://www.osp.ru/text/302/179534/).
5. Анил К. Джейн. Введение в искусственные нейронные сети. // Открытые системы. - № 4. - 1997. (http://www.osp.ru/text/302/179189/).
6. И. Шахнович. Нейронные сети в России – благодаря или вопреки? // ЭЛЕКТРОНИКА: Наука, Технология, Бизнес. - № 1. – 1999. (http://www.module.ru/files/papers-elestb0199.pdf).
7. Нейронные сети. Вводный курс. Сумской государственный университет. (http://neuronets.chat.ru/)



Двухслойный перцептрон для определения тональности текста

ЛЕКЦИЯ 11. ОСНОВЫ КОНЦЕПЦИИ ГЛУБИННОГО АНАЛИЗА ТЕКСТОВ (TEXT MINING)

1) Контент-анализ

Один из истоков концепции Text Mining – контент-анализ. Понятие контент-анализа, корни которого в психологии и социологии, не имеет однозначного определения:

- Контент-анализ - это методика объективного качественного и систематического изучения содержания средств коммуникации (Д. Джери, Дж. Джери).
- Контент-анализ - это систематическая числовая обработка, оценка и интерпретация формы и содержания информационного источника (Д. Мангейм, Р. Рич).
- Контент-анализ - это качественно-количественный метод изучения документов, которое характеризуется объективностью выводов и строгостью процедуры и состоит из квантификационной обработки текста с дальнейшей интерпретацией результатов (В. Иванов).
- Контент-анализ состоит из нахождения в тексте определенных содержательных понятий (единиц анализа), выявление частоты их встречаемости и соотношение с содержанием всего документа (Б. Краснов).
- Контент-анализ - это исследовательская техника для получения результатов путем анализа содержания текста о состоянии и свойствах социальной действительности (Э. Таршис).

Контент-анализ в рамках исследования электронных информационных массивов - относительно новое направление, которое предусматривает анализ множеств текстовых документов.

Принято распределение методологий контент-анализа на две области: качественную и количественную. Основа количественного контент-анализа - частота появления в документах определенных характеристик содержания. Качественный контент-анализ основан на самом факте присутствия или отсутствия в тексте одной или нескольких характеристик содержания.

2) Основные элементы концепции Text Mining

Технологии глубинного анализа текста Text Mining исторически предшествовала технология добычи данных, методология и подходы которой широко используются.

Важная задача технологии Text Mining связана с извлечением из текста его характерных элементов или свойств, которые могут использоваться как метаданные документа, ключевых слов, аннотаций. Другая важная задача состоит в отнесении документа к некоторым категориям

из заданной схемы их систематизации. Text Mining также обеспечивает новый уровень семантического поиска документов.

В соответствии с уже сформированной методологии к основным элементам Text Mining относятся: классификация (classification), кластеризация (clustering), построение семантических сетей, извлечение фактов, понятий (feature extraction), суммаризация (summarization), ответ на запросы (question answering), тематическое индексирование (thematic indexing) и поиск по ключевым словам (keyword searching). Также в некоторых случаях набор дополняют средства поддержки и создание таксономии (oftaxonomies) и тезаурусов (thesauri).

3) Классификация

При классификации текстов используются статистические корреляции для построения правил размещения документов в определенные категории. Задача классификации - это классическая задачу распознавания, где по некоторой контрольной выборке система относит новый объект к той или другой категории. Особенность же системы Text Mining заключается в том, что количество объектов и их атрибутов может быть очень большой, поэтому должны быть предусмотрены интеллектуальные механизмы оптимизации процесса классификации.

4) Кластеризация

Кластеризация базируется на признаках документов, которые использует лингвистические и математические методы без использования определенных категорий. Результат - таксономия или визуальная карта, которая обеспечивает эффективный охват больших объемов данных. Кластеризация в Text Mining рассматривается как процесс выделения компактных подгрупп объектов с близкими свойствами. Система должна самостоятельно найти признаки и разделить объекты по подгруппам. Кластеризация, как правило, передует классификации, поскольку разрешает определить группы объектов. Различают два основных типа кластеризации - иерархическую и бинарную.

5) Построение семантических сетей

Построение семантических сетей или анализ связей, которые определяют появление дескрипторов (ключевых фраз) в документе для обеспечения навигации.

6) Извлечение фактов

Извлечение фактов, предназначенное для получения некоторых фактов из текста с целью улучшения классификации, поиска и кластеризации.

7) Автоматическое реферирование

Автоматическое реферирование (Automatic Text Summarization) - это составление коротких изложений материалов, аннотаций или дайджестов, т.е. извлечения наиболее важных сведений из одного или нескольких документов и генерация на их основе лаконичных и информационно-насыщенных отчетов.

Существует множество путей решения задач, которые довольно четко подразделяются на два направлений - квазиреферирование и краткое изложение содержания первичных документов. Квазиреферирование основано на экстрагировании фрагментов документов - выделении наиболее информативных фраз и формировании из них квазирефератов.

Краткое изложение исходного материала основывается на выделении из текстов с помощью методов искусственного интеллекта и специальных информационных языков наиболее важной информации и порождении новых текстов, содержательно обобщающих первичные документы.

Семантические методы формирования рефератов-изложений допускают два основных подхода: метод синтаксического разбора предложений, и методы, базирующиеся на понимании естественного языка. Этот подход основывается на системах искусственного интеллекта, в которых также на этапе анализа выполняется синтаксический разбор текста, но синтаксические деревья не порождаются.

В рамках квазиреферирования выделяют три основных направления, применяемых совместно в современных системах:

- статистические методы, основанные на оценке информативности разных элементов текста по частоте встречаемости, которая служит основным критерием информативности слов, предложений или фраз;
- позиционные методы, которые опираются на предположение о том, что информативность элемента текста есть зависимым от его позиции в документе;
- индикаторные методы, основанные на оценке элементов текста, исходя из наличия в них специальных слов и словосочетаний - маркеров важности, что характеризуют их содержательную значимость.

Определение веса фрагментов (предложений или абзацев) исходного текста выполняется по алгоритмам, которые стали уже традиционными. Общий вес текстового блока на этом этапе определяется по формуле:

$$Weight = Location + KeyPhrase + StatTerm$$

Коэффициент Location определяется расположением блока в исходном тексте и зависит от того, где появляется данный фрагмент - в начале, в середине или в конце, а также используется ли он в ключевых разделах текста, например, в выводе.

Ключевые фразы (KeyPhrase) представляют собой конструкции-маркеры, которые резюмируют, типа "в заключение", "в данной статье", "в результате анализа" и т.п. Весовой коэффициент ключевой фразы может зависеть также от оценочного термина, например, "отличный".

Статистический вес текстового блока (StatTerm) вычисляется как нормированная по длине блока сумма весов входящих в него строк - слов и словосочетаний.

8) Поисковые образы документов

На основе методов автоматического реферирования возможно формирование поисковых образов документов. По автоматически построенным аннотациям больших текстов (поисковым образам документов) проводится поиск, который характеризуется высокой точностью (естественно, за счет полноты). Т.е. вместо поиска по полным текстам в некоторых случаях может оказаться целесообразным поиск по специально созданным аннотациям - поисковых образах документов.

В этом случае аннотированные тексты рассматри-ваются как поисковые образы документов (ПОД). Хотя ПОД часто для больших документов оказывается образованием, лишь отдаленно напоминающим исходный текст и не всегда оказывается воспринимаемым человеком, но за счет содержания наиболее весомых ключевых слов и фраз, он может приводить к вполне адекватным результатам при полнотекстовом поиске.

9) Особенности реализации систем с элементами Text Mining

Рассматриваются особенности таких систем:

- Intelligent Miner for Text (IBM)
- PolyAnalyst, WebAnalyst (Мегапьютер Інтеллидженс)
- Text Miner (SAS)
- SemioMap (Semio Corp.)
- Oracle Text (Oracle)
- Knowledge Server (Autonomy)
- RetrievalWare (Convera)
- Galaktika-ZOOM (корпорация "Галактика")
- InfoStream (Информационный центр "ЭЛВИСТИ")

Литература

Публикации Д.В. Ландэ по тематике лекции:

1. Ландэ Д.В. Добыча знаний. // Телеком. – 2004. - № 1-2. - С. 36-42. (http://dwl.visti.net/art/cz/, http://poiskbook.kiev.ua/cz.html).

2. Ландэ Д.В, Литвин А.Б. Феномены современных информационных потоков // Сети и бизнес. -2001. - № 1. - С. 14-21.(http://dwl.visti.net/art/content/).
3. Ландэ Д.В. Основы интеграции информационных потоков. – К.: Инжиниринг, 2006. – 240 с. (http://dwl.visti.net/art/monogr-osnov/spusk3.pdf)
4. Ландэ Д.В. Поиск знаний в Internet. –М.: Диалектика-Вильямс, 2005. (http://poiskbook.kiev.ua).

Публикации других авторов:

1. Хан Удо, Мани Индерджиет. Системы автоматического реферирования. // Открытые системы. -2000. - № 12. (http://www.osp.ru/os/2000/12/067.htm).
2. Michael W. Berry. Survey of Text Mining. Clustering, Classification, and Retrieval. - Springer-Verlag, 2004. - 244 p.
3. Chakrabarti Soumen. Mining the web. Discovery knowledge from hypertext data. - Publisher: Morgan Kaufmann, 2002. - 344 p.
4. Anthony Scime. Web mining: application and techniques. - Idea Group Publishing, 2005. - 427 p.

ЛЕКЦИЯ 12. КОНЦЕПЦИЯ И РЕАЛИЗАЦИЯ ТЕХНОЛОГИИ WIKI

Одним из видов источников информации, содержащих фактографические данные, являются сетевые энциклопедии, многие из которых сегодня имеют глобальный распределенный характер как с точки зрения потребления информации, так и с точки зрения их информационного наполнения. Многие сетевые информационно-поисковые системы, в частности, Yahoo! индексируют такие ресурсы и предъявляют их пользователям в первую очередь. В настоящее время самой крупной сетевой энциклопедией является Википедия, которая базируется на технологии Wiki. Остановимся подробнее на этом феномене. Основная идея Wiki-технологии состоит в обеспечении возможности коллективной работы с документами - любой документ из электронной библиотеки подлежит редакции любым пользователем.

1) Идеологические предпосылки Wiki

Идея “глобальной базы знаний” или “коллективного разума”, продукта совместного творчества пользователей Интернет.

Бесплатные онлайн-библиотеки.

Общедоступные энциклопедии.

2) Коллективное документирование

Основная идея Wiki -технологии состоит в обеспечении возможности коллективной работы с документами - любой документ из электронной библиотеки подлежит редакции любым пользователем.

Отмечается близость концепций Wiki и Open Source (открытый код).

Wiki -системы - это Web-сайты, работающие по принципу Вики, т.е. которые можно не только читать, но и изменять в режиме онлайн.

Сегодня большинство Wiki -систем размещаются на публичных серверах и являются объектом информационного вклада любого посетителя.

3) Сетевые энциклопедии

В свободном же доступе находятся как правило устаревшие или ограниченные версии, например, «Британника» 1911 года (http://1911encyclopedia.org) или Microsoft Encarta (4,5 тысячи статей из 60 тысяч). В России Яndex дает доступ к Большой советской энциклопедии 1978 года и Словарю Брокгауза и Ефрона 1907 года. Мегаэнциклопедия Кирилла и Мефодия (mega.km.ru) практически представляет собой электронную версию Современного Энциклопедического словаря 1997 года. Достаточно актуальная энциклопедия «Кругосвет» (krugosvet.ru) содержит лишь около 10 тысяч статей.

Сегодня определенной критической отметки достигли онлайновые энциклопедии, к которым доброжелательно относятся даже в научном мире. К таким проектам относится, прежде всего, Википедия.

4) Википедия

Проект Википедия ориентирован на создание онлайновой энциклопедии, написанной самими пользователями. Каждый посетитель сайта энциклопедии может внести свой посильный вклад: подправить статью, добавить или удалить информацию. Основной принцип Википедии - полнейшая демократия, поэтому начинающий автор может участвовать в проекте наряду с опытными экспертами.

Википедия - это полноценная онлайновая энциклопедия с 500 тыс. англоязычных статей, 1 млн. статей на 186 других языках (прежде всего немецком, японском и французском) и около 50 тыс. авторов и редакторов. Википедия, за короткий срок ставшая одним из популярнейших справочных ресурсов Интернета,

Статьи в Википедии написаны достаточно грамотно и часто даже превосходят по точности и полноте изложения статьи из традиционных энциклопедий, превосходя последние по оперативности, потому что:

- в системе продуманы простые и четкие правила составления и редакции статей;
- в системе предусмотрено «общественное» администрирование со своей иерархией администраторов, которые назначаются исходя из их вклада в развитие системы;
- вандалов, злоумышленников (или, как их называют в Википедии, «троллей») в действительности в десятки раз меньше, чем участников проекта, вносящих позитивный вклад.

5) Технология Wiki

Википедия базируется на Web-технологии Вики. Первая вики-среда была изобретена Каннингемом для веб-узла Pattern Languages Community с целью упростить совместное создание и ведения программных документов (свободно доступное программное обеспечение Вики на PHP доступно по адресу http://wikipedia.sourceforge.net, а первая его реализация доступна по адресу http://c2.com/cgi/wiki). Все страницы вики-сайта представляют собой статьи, содержимое которых представляет из себя текст, в котором можно использовать простую вики-разметку или тэги HTML.

Возможность редактировать содержимое вики-сайта любым посетителем, с одной стороны, позволяет без труда накапливать и систематизировать информацию, но, с другой стороны, создаёт обширное поле для внесения ошибок и вандализма.

Основные идеи, реализуемые Wiki -технологией:

- возможность редактирования Wiki -статей определенным кругом пользователей;
- хранение всех версий Wiki -статей с момента их создания;
- быстрая и простая генерация гиперссылок между документами, а также поддержка целостности гиперссылок;
- упрощение процесса публикации текста.

Корпоративные решения

Технология Вики используется не только в онлайновых энциклопедиях. Крупные компании как Motorola и The New York Times используют Wiki во всех рабочих процессах коллективного ведения проектной документации.

То, что сегодня разрабатывается компанией Microsoft для портала-сервера Sharepoint также можно квалифицировать как Wiki-технологию. Общее использование файлов далеко уходит от прежнего «разделения документов» (sharing). Происходит переход к совместным библиотекам, откуда документы можно легко достать, поработать с ними и положить на место более удобным, чем раньше, образом.

6) Язык разметки Wiki

Страницы Wiki-сайта представляют собой статьи, содержимое которых - это обычный текст, в котором можно использовать теги HTML или особую вики-разметку, более удобную для тек-

стовых документов, чем HTML. Воспользовавшись ссылкой или кнопкой, любой посетитель вики-сайта может отредактировать и сохранить измененный вариант текста любой существующей страницы или создать новую. Процедура публикации текста в Википедии сведена к двум кнопкам - редактировать и сохранить.

Определенная часть статей в Википедии представляет собой созданные автоматически "заготовки". Когда любой автор отмечает в тексте термин или выражение как ссылку на несуществующую статью, в Википедии автоматически генерируется новая статья-шаблон, содержащая текст: "статья еще не написана, можете ее написать". Авторы, заходящие по этой ссылке, расширяют содержание «пустышки».

Для публикации, например, математических формул в Википедии используется разметка TeX -языка программирования специального назначения, который широко применяется при подготовке текстов научно-технического характера, создателем которого был Д. Кнут.

В зависимости от желания автора и сложности формул, записываемых в текстах статей, генерируются либо изображения формул в формате PNG, либо простой код HTML.

История правок всех Wiki-статей хранится в базе данных, любая редакция статьи может быть вызвана на экран и сохранена, как последняя.

В Википедии имеется автоматизированная генерация и поддержание целостности гиперссылок между документами на всем сайте. Проблемы с неразрешенными ссылками в Википедии не существует. В Wiki -среде, естественно, можно использовать и таблицы, их форматирование также упрощено, по сравнению с HTML, однако, любители последнего могут воспользоваться интерактивным конвертором HTML-таблиц в синтаксис системы Wiki.

Физически Википедия размещается на 25 серверах, расположенных в США. Серверы работают под управлением Linux (RedHat 9 и Fedora Core). На двух хранится база данных (MySQL), три используются для кэширования запросов (Squid), пятнадцать работают как веб-серверы (Apache). Остальные являются почтовыми и/или DNS-серверами.

7) Современное состояние

Благодаря успеху Википедия у проекта появились новые ответвления. Сейчас на сайте Wikimedia Foundation размещаются проекты Вики-словарей, Вики-цитатник, Вики-учебники, Вики-программные коды и ряд других. Существует агентство новостей Wikinews (Wiki-новости).

На сегодняшний день англоязычная Википедия самая большая в мире - свыше 1,4 млн. статей. Вторая по величине Википедия - немецкая, она содержит более 475 тысяч статей. Третья по величине Википедия - французская, у нее сейчас 371 тысячу статей.

Википедия представляет собой альтернативный подход к созданию семантического Web для достижения той же цели – получения знаний. Поиск в Википедии даже по обычным ключевым словам приводит к получению содержательных материалов, содержащих сконцентрированные знания.

За время своего развития проект Википедии, несмотря на опасения, связанные с непрофессиональностью авторов, возможным вандализмом, спонтанностью создания отдельных статей, позволил создать достаточно качественный продукт – полную и объективную, свободно доступную всем многоязычную энциклопедию.

Литература

Публикации Д.В. Ландэ тематике лекции:

1. Ландэ Д.В. За знаниями - к Википедии. // Телеком. – 2005. - №№. 9, 11. (http://dwl.visti.net/art/wiki/)
2. Ландэ Д.В. Основы интеграции информационных потоков. – К.: Инжиниринг, 2006. – 240 с. (http://dwl.visti.net/art/monogr-osnov/spusk3.pdf)

Публикации других авторов:

1. Русский сайт Википедии (ru.wikipedia.org)
2. А. Птица. Планета Wiki. Домашний ПК. -№ 12. – 2005 (http://itc.kiev.ua/article.phtml?ID=22663&IDw=10)

3. А. Черников. «Википедия», или Магия Open Approach. Компьютерное обозрение. 16 декабря 2005 (http://itc.kiev.ua/article.phtml?ID=22981)
4. А. Москалюк. Web 2.0: паутина как платформа. Компьютерное обозрение. 27 октября 2004 (http://itc.kiev.ua/article.phtml?ID=18767)
5. Daniel Terdiman. Wiki Becomes a Way of Life. Wired. 2005-03-08 (http://www.wired.com/news/culture/1,66814-0.html)

ЛЕКЦИЯ 13. ОСНОВНЫЕ СВЕДЕНИЯ О КОНЦЕПЦИИ СЕМАНТИЧЕСКОГО WEB

1) Проблемы традиционного Web:

- рост объемов информационного наполнения;
- представление информации, которое ориентированно преимущественно на людей;
- проблема нахождения необходимой пользователю информации;
- невозможность выделить смысл сообщений в автоматизированном режиме.

2) Основные идеи Семантического Web

Возможность интегрировать в Интернет объекты реального мира благодаря унификации обмена данными.

Организация такого представления данных в сети, чтобы допускалась не только их визуализация, но и их эффективная автоматическая обработка программами разных производителей.

Создание непрерывного информационного поля, превращение его в систему семантического уровня.

Семантический Web представляет собой расширение существующей сети Интернет, в котором информация представляется в четком и определенном смысловом значении, дающем возможность людям и компьютерам работать с более высокой степенью взаимопонимания и согласованности.

3) Структура семантического Web

В процессе реализации концепции Семантического Web получили широкое развитие синтаксические методы представления информации языковыми средствами XML и его дополнений, предназначенных для описания типовых свойств элементов XML-документов, их структуры и семантики: рекомендации W3C, регламентирующие DTD (Document Type Definition), XML Schema, XQuery (язык запросов к базам XML-данных) и т.д. К языкам представления данных относятся также Средства Описания Ресурсов RDF (Resource Description Framework). Существует также ряд других форматов, однако XML и RDF предоставляют больше возможностей, потому они уже обладают статусом рекомендаций W3C.

Другая ветвь Семантического Web связана с направлениями, близкими к области искусственного интеллекта, и названа онтологическим подходом. Этот подход включает в себя средства аннотирования документов, которыми могли бы воспользоваться компьютерные программы - Web-сервисы и агенты при обработке сложных пользовательских запросов. Модели предметных областей в терминологии Семантического Web называются онтологиями. 10 февраля 2004 года консорциумом W3C была утверждена и опубликована спецификация языка сетевых онтологий OWL (Web Ontology Language).

Две ветви Семантического Web используют три ключевых языка (соответственно, технологий):

- спецификация XML, позволяющая определить синтаксис и структуру документов;
- механизм описания ресурсов RDF, обеспечивающий модель кодирования для значений понятий, определенных в онтологиях.
- язык онтологий OWL, позволяющий определять понятия и отношения между ними.

Семантический Web использует также и другие языки, технологии и концепции, в частности, универсальные идентификаторы ресурсов, цифровые подписи, системы логического вывода и т. д.

При этом самый нижний уровень Семантического Web — это Universal Resource Identifier (URI), унифицированный идентификатор, определяющий способ записи адреса произвольного ресурса.

Отдельный уровень в концепции Семантического Web ориентирован на работу с цифровой подписью, которая необходима, чтобы клиенты могли определять степень достоверности данных.

4) XML – синтаксическая основа Семантического Web

Исходная версия XML, разработанная в консорциуме W3C под руководством Джона Босака, была опубликована в феврале 1998 года и с тех пор развилась до уровня метаязыка, на базе которого определяются сотни новых предметно-ориентированных языков (к примеру, MathML, XLink, SMIL, XSL и др.)

В отличие от HTML, XML предназначен для разметки документов произвольной структуры. Универсальный синтаксис XML обусловил появление ряда технологий, таких как XSL и XPath, предназначенные для работы с древовидной структурой документов; XML Schema – стандарт описания конкретных языков разметки, использующий синтаксис XML; XLink и XPointer – средства связи распределенных блоков информации в один общий документ; XQuery – язык запросов к XML-данным

Формат любого тега XML прост: *<идентификатор> содержание </идентификатор>*.

Поскольку в XML не существует фиксированного словаря тегов, то они могут определяться независимо для каждой программы. В XML это было изначально предусмотрено с помощью определения типа документа DTD (Document Type Definitions), накладывающего ограничения на используемые теги и задающего грамматику, которая указывает допустимые комбинации и вложения имен тегов, имен атрибутов и т.д.

Вместе с тем, языку DTD присущи два серьезных недостатка - ограниченность описания типов данных и синтаксис, отличный от XML. Поэтому в настоящее время консорциум W3C настоятельно рекомендует заменять использование DTD новым стандартом - XML-схем (XML Schema), который был утвержден в 2001 году (http://www.w3.org/TR/xmlschema-formal/).

5) Средства описания ресурсов RDF

RDF - язык формального описания содержания сетевых ресурсов, который согласно архитектуре Семантического Web представляет собой связующее звено между XML-документами и средствами, обеспечивающими поиск и навигацию на основе логических утверждений.
Принцип построения отношений между сетевыми ресурсами в спецификации RDF предусматривает наличие трех компонент - объекта, атрибута и значения (аналогичных классической схеме "подлежащее - сказуемое - дополнение").

Базовый строительный блок в RDF - триплет "объект - атрибут - значение" часто записывают в виде A(O,V), где O – объект (ресурс), A - атрибут (свойство) со значением (субъектом) V. RDF позволяет менять местами объекты и значения. Благодаря тому, что RDF использует URI-идентификаторы для кодирования информации в документе, обеспечивается возможная привязка понятия к единому определению, которое можно найти в Сети.

Модель данных RDF сама по себе является всего лишь синтаксической основой - для того чтобы описание обрело смысл, необходимо воспользоваться словарями терминов и понятий, которые задаются с помощью технологии – RDF схема (Schema), играющей для RDF такую же роль, что и схема для XML.

RDF — это самый низкоуровневый из существующих языков описания метаданных, поскольку оперирует лишь понятиями связей примитивных сущностей, например, «объект А владеет субъектом Б».

6) Онтологии

В структуре Семантического Web предусмотрены и более эффективные специальные средства описания содержания, чем RDF. Онтологии - это базы знаний, которые включают в себя сведения, необходимые для отождествления новых понятий с уже известными, для определения принадлежности используемых терминов к той или иной предметной области и, в результате, для приведения любых понятий к виду, пригодному для восприятия программами - электронными агентами.Отличия онтологий от XML:

- Онтология отличается тем, что это представление знания, а не формат документов.
- Одним из преимуществ онтологий будет доступность инструментов, которые обеспечат универсальную поддержку семантики, которая не является специфической для определенной предметной области.

Предполагается, что «интеллектуальные» приложения смогут использовать онтологии, чтобы получать в результате поиска информацию со связанной с ней структурой знаний и правилами вывода.

Разработан и стандартизирован язык описания структурированных онтологий OWL. В рамках OWL онтология – это совокупность утверждений, задающих отношения между понятиями и определяющих логические правила для рассуждений о них.

Онтология может включать описания классов, свойств и их примеры. OWL может использоваться, чтобы явно представлять значения терминов и отношения между этими терминами в словарях. OWL имеет больше средств для выражения значения и семантики, чем XML, RDF, и RDF-S, и, таким образом, OWL идет дальше этих языков в способности представить поддающийся машинной обработке контент Сети.

7) Знания

Семантический Web предполагает создание системы с элементами "искусственного интеллекта", которая бы позволила специальным приложениям качественно искать в Интернет необходимую информацию, а также обмениваться информацией друг с другом. При этом именно язык онтологий OWL выступил решающей компонентой интеллектуализации, базисом для построения семантических сетей.

Представлениям знаний в Семантическом Web присущи универсальные выразительные возможности, синтаксическая и семантическая интероперабельность, которая реализуется, например, в онтологиях путем установлением соответствия между используемыми терминами.

8) Агенты

Под агентами понимаются программы, работающие без непосредственного управления со стороны человека для достижения поставленных перед ней целей. Обычно агенты собирают, фильтруют и обрабатывают информацию, найденную в Сети, иногда путем взаимодействия с другими агентами. В соответствии с документами W3C, Семантический Web заработает в полную силу тогда, когда люди создадут множество программ, которые, знакомясь с содержимым Сети из различных источников, смогут обрабатывать полученную информацию и обмениваться результатами с другими программами.

Основные принципы функционирования агентов:

- агент не имеет полной информации, необходимой для решения поставленной задачи;
- обрабатываемые данные распределены в сети;
- вычисления выполняются агентами асинхронно;
- взаимодействие агентов друг с другом и с человеком происходит на высоком семантическом уровне;
- отсутствует глобальный контроль за деятельностью всей системы агентов.

Литература

Публикации Д.В. Ландэ по тематике лекции:

1. Ландэ Д.В. Семантический веб: воплощение идеи. // Телеком. – 2005. - №6, 2005. - С. 60-65. (http://dwl.visti.net/art/sw/)

2. Ландэ Д.В. На границе стихий. // CHIP/Украина. – 2003. - №5, - С. 72-77 (http://dwl.kiev.ua/art/xml/).
3. Ландэ Д.В. Основы интеграции информационных потоков. – К.: Инжиниринг, 2006. – 240 с. (http://dwl.visti.net/art/monogr-osnov/spusk3.pdf)

Публикации других авторов:

1. Tim Berners-Lee, James Hendler, Ora Lassila. The Semantic Web. Scientific American, 2001 (http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21)
2. The Semantic Web Portal (http://www.w3.org/2001/sw/).
3. Eric Miller, Ralph Swick, Dan Brickley, Brian McBride, Jim Hendler, Guus Schreiber, Dan Connolly. Semantic Web. W3C (MIT, ERCIM, Keio). - 2001. (http://www.w3.org/2001/sw/)
4. Quin Liam. Extensible Markup Language (XML). – 2003. (http://www.w3.org/XML/)

ЛЕКЦИЯ 14. ОСНОВНЫЕ ЗАКОНОМЕРНОСТИ РАЗВИТИЯ ИНФОРМАЦИОННОГО ПРОСТРАНСТВА

1) Правило Парето

Анаизируя общественные процессы, Парето рассматривал социальную среду как пирамиду, наверху которой находятся немногие люди, составляющие элиту. В результате кропотливых исследований ученый сформулировал математическую зависимость между величиной дохода и количеством получающих его лиц. Ученый в 1906 году установил, что 80 процентов земли в Италии принадлежит лишь 20 процентам ее жителей. Парето пришел к выводу, что параметры полученного им распределения примерно одинаковы и не различаются принципиально в разных странах и в разное время.

Распределение доходов по Парето описывается уравнением $N = A / X^{p+1}$, где X – величина дохода, N - численность людей с доходом, равным или выше X, A и p - коэффициенты уравнения. В математической статистике это распределение получило имя Парето, при этом естественные ограничения на коэффициенты: $X \geq 1$, $p > 0$. Распределение Парето обладает свойством устойчивости, т.е. сумма двух случайных переменных, имеющих распределение Парето, также будет иметь это распределение.

Замеченное правило применимо и в очень многих областях и сформулировал правило, называемое "Закон Парето" или "Принцип 80/20". Например, при информационном поиске достаточно определить 20% необходимых ключевых слов, после чего найти 80% требуемых документов, а затем расширить поиск или воспользоваться опцией "найти похожие" для полного решения задачи. Еще один пример: 80% посещений Web-сайта приходится лишь на 20% его Web-страниц.

При реализации систем массового обслуживания, в том числе и поисковых систем, необходимо учитывать то, что наиболее сложным функциональным возможностям системы, на реализацию которых ушло 80 и более процентов трудозатрат будут использоваться не более, чем 20% пользователей данной системы.

2) О переходе количества в качество

Если система достигла 99% своей идеальной функциональности, то дальнейшие попытки ее совершенствования ведут, в лучшем случае, к повышению качества сопровождения реализованных уже функций, и, если изобразить график, отмечая по оси абсцисс затраченные ресурсы на развитие системы, а по оси ординат - уровень функциональности, то график будет иметь вид кривой, у которой в начале наблюдается резкий подъем, и которая стабилизируется. В то же время, реализация новых подходов приводит к появлению новых, даже не предполагаемых ранее показателей.

В качестве примера этой закономерности можно привести развитие сети Internet, которая до начала 90-х годов прошлого века рассматривалась, прежде всего, как компьютерная сеть передачи данных, а уж затем, как хранилище информационных ресурсов. Несмотря на то, что существовали такие информационные службы, как Usenet, Ftp, Gopher, до 90-х годов Сеть ре-

шала свои главные задачи, обеспечивая электронную связь между научными, общественными, государственными организациями и частными лицами. К этому времени Интернет существовал уже свыше 15-ти лет и стабилизировалась в своем развитии, в частности, по числу абонентов. Феномен появления и развития Web-технологий привел к тому, что за следующие 10 лет сеть Интернет стала крупнейшим информационным ресурсом в мире, число абонентов которой превысило миллиард человек.

3) Законы Зипфа

При статистическом описании распределения слов по частоте их употребления в тексте (как, впрочем, и в документальных потоках) используются так называемые ранговые распределения (ранг - это, например, порядковый номер слова в списке, где все слова упорядочены по возрастанию относительных частот).

Джордж Зипф экспериментально показал, что распределение слов естественного языка подчиняется закону, который можно сформулировать следующим образом. Если к какому-либо достаточно большому тексту составить список всех встретившихся в нем слов, а затем проранжировать эти слова в порядке убывания частоты их встречаемости в тексте, то для любого слова произведение его ранга и частоты встречаемости будет величиной постоянной: $\boldsymbol{f * r = c}$, где $\boldsymbol{f}$ - частота встречаемости слова в тексте; $\boldsymbol{r}$ - ранг слова в списке; $\boldsymbol{c}$ - эмпирическая постоянная величина. Для русского и украинского языков коэффициенты Зипфа составляю приблизительно 0,06-0,07.

Зипф сформулировал еще одну закономерность, состоящую в том, что частота и количество слов, входящих в текст с данной частотой, также связаны подобным соотношением.

Известный математик Бенуа Мандлеброт математическим путем пришел к аналогичной первому закону Ципфа зависимости $\boldsymbol{f*r^e = c}$, где $\boldsymbol{e}$ - близкая к единице переменная величина, которая может изменяться в зависимости от свойств текста и языка.

Законам Зипфа удовлетворяют не только слова из одного текста, но и практически все объекты современного информационного пространства.

4) Закономерность Брэдфорда

Основной смысл закономерности С. Брэдфорда заключается в следующем: если научные журналы расположить в порядке убывания числа помещенных в них статей по конкретному предмету, то полученный список можно разбить на три зоны таким образом, чтобы количество статей в каждой зоне по заданному предмету было одинаковым. Эти три зоны составляли: ядро - профильные журналы, непосредственно посвященные рассматриваемой тематике, журналы, частично посвященные заданной области, и журналы, тематика которых весьма далека от рассматриваемого предмета. С. Брэдфорд установил, что количество журналов в третьей зоне будет примерно во столько раз больше, чем во второй зоне, во сколько раз число наименований во второй зоне больше, чем в ядре, т.е. $P3 : P2 = P2 : P1 = N$, где Р1 - число журналов в 1-й зоне, $P2$ - во 2-й, $P3$ - число журналов в 3-й зоне.

Закономерность Брэдфорда изначально рассматривалась как специфический случай распределения Зипфа для системы периодических изданий по науке и технике. Исходя из реалий развития сети Интернет, ее можно рассматривать как закономерность, относящуюся к ранговому распределению Web-сайтов, относительно вхождения в них Web-страниц, релевантных некоторой области знаний.

5) Закон Хипса

В компьютерной лингвистике эмпирический закон Хипса связывает объем документа с объемом словаря уникальных слов, которые входят в этот документ. Казалось бы, словарь уникальных слов должен насыщаться, а его объем стабилизироваться при увеличении объемов текста. Оказывается это не так! В соответствии закону Хипса, эти значения связаны соотношением:

$$v(n) = Kn^{\beta},$$

где v – это объем словаря уникальных слов, составленный из текста, который состоит из n уникальных слов. K и β – обусловленные эмпирически параметры. Для европейских языков K принимает значение от 10 до100, а β - от 0.4 до 0.6.

Закон Хипса справедлив не только для уникальных слов, но и для многих других информационных объектов, описываемых не экспоненциальной, а степенной зависимостью.

6) Прогноз Мура и информационная сфера

Эта закономерность, которая родилась как прогноз развития технологии микросхем, но все шире вторгается во все сферы жизни. В 1965 году Гордон Мур предсказал, что плотность транзисторов в интегральных схемах и, соответственно, производительность микропроцессоров будут удваиваться каждый год. В течение трех последних десятилетий этот прогноз, названный "законом Мура", более или менее выполнялся, хотя достаточно быстро был скорректирован - удвоение должно происходить каждые два года.

Сегодня прогноз Мура распространяется на все большее количество областей. Сегодняшнее расширение Internet, стремительный рост объемов пересылаемых данных, развитие электронной коммерции и беспроводной связи, а также внедрение цифровых технологий в бытовую технику, можно рассматривать как следствие этого закона Мура. Было замечено, что рост документальной информации, вполне подчиняясь закону Мура, также носит экспоненциальный характер, а именно кривая роста числа документов может быть описана уравнением вида $\boldsymbol{y = Ae^{kt}}$, где y – количество документов, t – время (г); A – количество документов начале отсчета, k – коэффициент.

Развитие коммуникационных возможностей приводит к росту количества доступной информации, в частности в Интернет. С другой стороны, увеличение объемов доступного контента способствует росту инновационной деятельности, все больше знаний, необходимых для исследовательских работ, публикуется в Сети, тем самым, способствуя технологическому прогрессу, на котором основывается прогноз Мура.

Литература

Публикации Д.В. Ландэ по тематике лекции:

1. Ландэ Д.В. Поиск знаний в Internet. –М.: Диалектика-Вильямс, 2005. (http://poiskbook.kiev.ua)
2. Ландэ Д.В, Литвин А.Б. Феномены современных информационных потоков // Сети и бизнес. -2001. - № 1. - С. 14-21.(http://dwl.visti.net/art/content/)

Публикации других авторов:

1. Чурсин Н.Н. Популярная информатика. К.: Техника, 1982.
2. Кларк Д. Закон Мура останется в силе // Ведомости. – 2003. - № 11. (http://www.silicontaiga.ru/home.asp?artId=2066).
4. Иванов С.А. Устойчивые закономерности мировой системы научной коммуникации. // Научно-техническая информация. Сер. 2. Вып. 1. - 2003. - С. 1-7.
5. Иванов С.А. Ранговые распределения в информатике. // Научно-техническая информация. Сер. 2. Вып. 12. - 1985. - С. 14-19.
6. Алексеев Н.Г**.** Применение закона Бредфорда при комплектовании фонда научной библиотеки. // Тезисы докладов конференциии "Библиотечное дело-1996" (http://libconfs.narod.ru/1996/4s/4s_p1.html)

КОНТРОЛЬНЫЕ ВОПРОСЫ ПО КУРСУ

1. Общая информация об Интернет
2. Гипертекст и WWW
3. Топология Web-пространства
4. "Скрытый" Web
5. Характеристики ИПС, метрики РОМИП

6. Информационно-поисковые языки
7. Ранжирование
8. Булева модель поиска
9. Векторно-пространственная модель поиска
10. Вероятностная модель поиска
11. Клатеризация. Обзор методов
12. Статическая и динамическая составляющие Web-пространства
13. Синдикация новостной информации
14. RSS-формат
15. Модели динамики информационных потоков
16. Фрактальные свойства информационных потоков. Параметр Херста.
17. Вики (корпоративная технология и Википедия)
18. Разметка документов в Вики
19. Нейронные сети (основы), перцептроны
20. Основные компоненты семантического Web
21. Text Mining общее представление
22. Автоматическое реферирование
23. Правило Парето и некоторые его следствия
24. Законы Зипфа
25. Закономерность Брэдфорда