С.А.ШУМСКИЙ
ФИАН им.П.Н.Лебедева, ООО «НейрОК», Москва
E-mail: shumsky@neurok.ru

САМООРГАНИЗУЮЩИЕСЯ СЕМАНТИЧЕСКИЕ СЕТИ

Аннотация

Интернет представляет собой гигантскую распределенную базу общедоступных данных. Однако существующие средства поиска и фильтрации информации не позволяют с достаточной эффективностью использовать имеющиеся в Сети знания. С другой стороны, бо́льшая часть компьютерных ресурсов – сотни миллионов персональных компьютеров – большую часть времени простаивает. В лекции рассматривается вариант создания самоорганизующейся распределенной семантической сети, состоящей из взаимодействующих друг с другом обучающихся поисковых агентов. Агенты постоянно добывают новую информацию из Сети, одновременно обучаясь обслуживать интересы своих хозяев. Агентская сеть способна решить многие актуальные задачи Интернет: своевременной индексации информации, направленного распространения персональной, новостной и коммерческой информации, а также автоматической организации «клубов по интересам».

S.A.SHUMSKY
Lebedev Physics Institute, «NeurOK» LLC, Moscow
E-mail: shumsky@neurok.ru

SELFORGANIZING SEMANTIC NETWORKS

Abstract

Internet is the largest known publicly accessible database. But existing search and filtering tools fail to keep pace with the growing amount of information. On the other hand hundreds of million of installed PC stay idle most of time. This paper presents a basic sketch of a plan of distributed self-organizing semantic network, comprising interacting self-learning search agents. The latter constantly filter information on behalf of their owners, adapting to their personal profiles. Such self-learning semantic network may facilitate: real-time information indexing, personal news delivery, targeted commercials, and emergent communities.

 УДК 004.032.26 (06) Нейронные сети

Введение. Информационный фазовый переход

В последние несколько лет мы являемся свидетелями радикальных изменений в информационной инфраструктуре общества. Процесс тотальной компьютеризации, катализатором которого были в 80-е годы персональные компьютеры (ПК), а в 90-е – Интернет, качественно изменил глобальный информационный ландшафт.

К началу нового века практически все накопленные человечеством знания оказались, во-первых, оцифрованы и, во-вторых, общедоступны через глобальную Сеть. Создана принципиально новая, программируемая, среда доступа к этой информации.

Произошел, если можно так выразиться, «информационный фазовый переход»: вместо разрозненных «озер» локальных баз данных появился глобальный «океан» всемирной Сети с простыми средствами доступа к нему со стороны массового пользователя.[1]

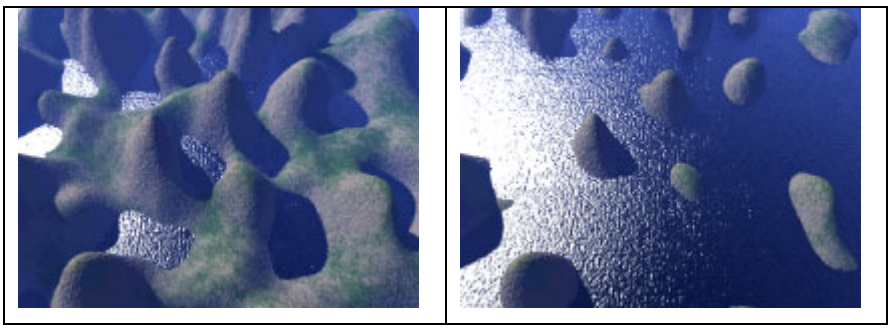

Рис. 1. Изменение топологии информационного пространства.

Этот океан имеет относительно тонкий «поверхностный слой», World Wide Web (WWW) или просто Web, состоящий из статичных html-страниц с постоянным адресом. Именно этот слой, содержащий в настоящее время порядка 2.5 миллиардов документов, индексируют (точнее,

[1] Подобный топологический фазовый переход наблюдается при сбивании масла (эмульсии воды в жире) из сливок (суспензии жира в воде).

стараются проиндексировать) всем известные поисковые серверы Интернет. Однако, есть еще гораздо более обширный «глубинный» Интернет, состоящий из подключенных к Сети баз данных общим объемом ~550 миллиардов документов. Эти документы выдаются по соответствующим запросам в виде динамических, формируемых на лету, html-страниц. Причем 95% этих баз данных находятся в общественном пользовании. Если на поверхности Океана лежит от 10 до 20 терабайт текстовой информации, то в его глубине сокрыто порядка 7500 терабайт. Информационные потоки в Океане еще более впечатляюще. Ежегодно люди посылают друг другу порядка 10^{12} электронных писем общим объемом свыше 10000 терабайт [1]. То есть «творческий потенциал» пользователей Сети сравним со всей накопленной в базах данных информацией.

Мы еще только привыкаем к жизни в этой новой для нас цифровой, программируемой среде, в которой можно создавать *любые* модели общения и доступа к информации. Для ее освоения наверняка потребуются новые средства поиска и фильтрации информации, новые средства навигации и общения людей как друг с другом, так и с миром программ. Интернет недалекого будущего, скорее всего, будет довольно сильно отличаться от существующего, использующего технологии, разработанные десятилетия назад для работы в локальных базах данных.

Многие, в том числе и автор данной статьи, считают неизбежным переход от централизованного к распределенному индексированию и поиску информации, поскольку ни один сервер не в состоянии в одиночку «объять необъятное». Более того, согласно приведенной выше статистике, поисковым серверам и сейчас доступна лишь «поверхность» Интернета. Основная часть информации находится в базах данных и предоставляется лишь через соответствующие *сервисы*. Однако, пока что локальные индексы баз данных разрознены, и поиск в Сети ассоциируется с поиском в статичных http-страницах Web, т.е. «на поверхности» Интернет. Глубинный Интернет пока недоступен для глобального поиска. Это относится как к корпоративным базам данных, так и к пользовательским ПК, содержащим, к слову, 55% мировой информации [2].

Идеальный поисковый сервис Сети должен, естественно, включать и информацию в корпоративных базах данных и открытую для публичного доступа информацию на персональных машинах. Для этого локальные

поисковые сервисы должны быть включены в единую поисковую сеть, в которой каждый *поисковый агент* знает других агентов, к которым можно автоматически перенаправить поиск в случае, если он сам не нашел ответа.

Точки входа в такую поисковую сеть, *персональные поисковые агенты*, к которым пользователи и обращаются со своими запросами, должны, в идеале, заранее знать кому эти запросы переадресовывать, а ссылки на наиболее важную информацию хранить непосредственно у себя в локальном поисковом индексе. Такие персональные поисковые агенты должны быть *обучаемыми* – уметь подстраиваться, с одной стороны, под интересы своих хозяев, а с другой – находить источники соответствующей информации в Сети.

Такие персональные обучающиеся агенты действительно смогут обеспечить необходимый информационный комфорт простым пользователям Сети, которым уже не нужно будет запоминать где можно найти ответы на те или иные вопросы, раздувая до невообразимых размеров дерево своих «закладок» по мере появления все новых сетевых сервисов. Общение с Сетью будет ассистировать их персональный помощник, предварительно находящий, прочитывающий и отбирающий для них наиболее интересную информацию.

В этой лекции мы рассмотрим, как может быть устроена подобная сеть взаимодействующих друг с другом персональных поисковых агентов, и какие функции смогут взять на себя эти электронные секретари в повседневной жизни пользователей Сети.

Ведь не поиском же единым жив человек. Новые, агентские технологии должны будут удовлетворять и другие потребности, возникающие в условиях глобального общежития, в частности – потребности в общении и обмене персональной информацией.

Действительно, как мы заметили выше, годовой объем информации, генерируемой пользователями Сети, сравним со всей накопленной человечеством информацией. Этой информацией люди обмениваются, в частности, в виде писем. Это – уровень персонального общения людей, уже знающих друг друга (*user-user*).

Другой полюс информационных потоков составляют средства массовой информации – *broadcasting*. Если ограничиться лишь печатной продукцией, которая сегодня практически вся доступна и через Web, то ее объем (~20 терабайт – см. **Рис. 2**) окажется пренебрежимо мал, по сравнению с объемом той же электронной почты (~10000 терабайт). Это и естественно, ведь broadcasting-информация производится относительно небольшой группой людей одной профессии – журналистами. Однако, именно их глазами мы вынуждены смотреть на мир, и именно их суждения влияют на нас в первую очередь, формируя общественное мнение по широкому кругу вопросов. Этот тип общения, «вещание» – broadcasting, был перенесен в Интернет многочисленными порталами, механически транслирующими бизнес-модель масс-медиа в новую, программируемую среду.

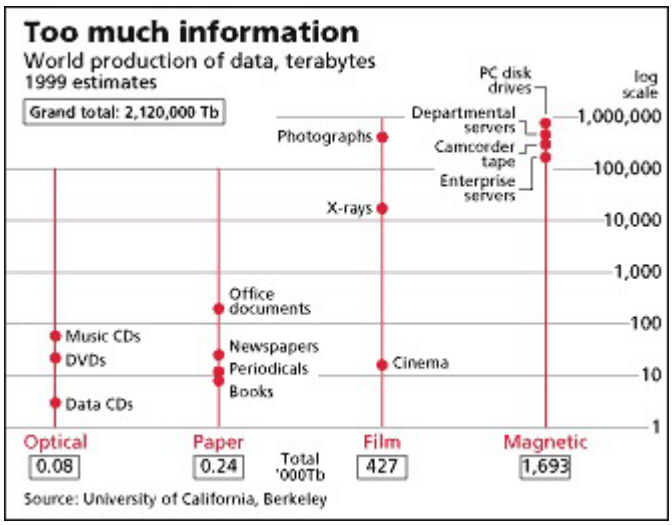

Рис. 2. Оценка сверху различных типов информации в мире (Economist Magazine, October, 2000 [2]).

Между тем, если вспомнить историю, Интернет первоначально культивировал альтернативную модель коллаборативного общения единомышленников – людей одной профессии или просто сходных интересов.

Сюда можно отнести, например, сеть обмена мнениями USENET, разбитую на десятки тысяч тематических подразделов. Годовой объем корреспонденции типа *user-community* в USENET оценивается в 70 терабайт [1]. Это всего лишь в три с небольшим раза больше, чем broadcasting, но на два порядка меньше, чем корреспонденция user-user.

Можно трактовать эти цифры по-разному. В частности, что люди не слишком-то и нуждаются в общении на уровне community. Узнать общечеловеческие новости по каналам broadcasting и обменяться информацией со своим ближайшим окружением средствами электронной почты – вот и все, что им по большому счету надо.

Альтернативная точка зрения состоит в том, что у людей нет другого выбора пока им не предоставят действительно удобного, «на кончиках пальцев», средства для социального общения в Сети. Таким инструментом как раз и может стать сеть персональных агентов, автоматически формирующих своему хозяину созвучное его интересам социальное окружение. Подобно тому, как Windows-интерфейс открыл массовому пользователю дорогу в мир компьютеров, а графический браузер Netscape – в мир WWW, агентский интерфейс способен, по нашему глубокому убеждению, кардинально изменить облик Сети.

Метафора программных агентов, активно помогающих пользователям в поисках действительно интересной им информации, очень естественна для жизни в открытом информационном мире. Она должна существенно потеснить ставший уже привычным *объектный* интерфейс Windows, позволяющий манипулировать с информацией как с обычными материальными объектами. Это максимально облегчает пользователю *самостоятельные* манипуляции с информацией. Агентский же интерфейс призван, напротив, *заменить* пользователя при выполнении большого числа рутинных операций, например, по предварительному просеиванию больших потоков информации.

Но для этого, как легко догадаться, агенты должны быть достаточно интеллектуальными – в какой-то мере *понимать* о чем идет речь в том или ином документе. Как научить агентов понимать *семантику* языка, уметь сравнивать содержимое различных документов и обучаться интересам пользователя? Об этом пойдет речь в следующем разделе.

Теоретические выводы мы подкрепим примером реальной коммерческой системы семантического анализа текстовых коллекций **Semantic Explorer**.

Далее мы рассмотрим вопрос формирования оптимальной структуры связей в поисковой сети таких семантических агентов. Мы покажем, что широко используемый в нейрокомпьютинге метод настройки связей (обучение сетей по методу обратного распространения ошибки – backpropagation) в некотором смысле эквивалентен экономическим отношениям между агентами, основанным на денежном обмене.

Наконец, мы рассмотрим некоторые социальные последствия появления в сети агентского сообщества – какие новые возможности появятся у пользователей, заведи они себе таких агентов.

Семантический анализ контента

В этом разделе мы сосредоточимся на проблеме обучения поисковых агентов распознаванию семантических образов. Под *семантическим* образом документа мы будем понимать его тематическую направленность – то, *о чем* говорится в данном документе. Это чрезвычайно сжатое представление, по нему невозможно восстановить не только конкретное содержание документа, но даже то, какими именно словами ведется изложение.

Соответственно, семантические поиск и фильтрация информации основаны не на ключевых словах, а, скорее, на ключевых понятиях. Если лексическое представление документа представляет из себя спектр составляющих его слов, точнее принимаемых во внимание базовых *терминов*, то семантическое представление – вектор присутствия в нем базовых *семантических категорий*. Причем, если характерное число слов в архивных коллекциях документов колеблется в диапазоне от 10^4 до 10^6, в зависимости от их объема и широты охвата, то число различаемых людьми тематических рубрик обычно на несколько порядков меньше – соответственно от 10^2 до 10^3. [2]

[2] Если же в качестве такой коллекции представить себе Интернет в целом, со всеми использующимися в нем языками, то эти цифры возрастут еще на несколько

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

Выработка способности семантического представления текстовой информации, соответственно, сопровождается обучением языку, точнее, конечно, лишь некоторым его аспектам – закономерностям совместного употребления слов в документах. Но именно эти закономерности и определяют во многом значения слов. Действительно, смысл слова зависит, естественно, не от его написания, а от его употребления, т.е. определяется совокупностью всех тех его комбинаций с другими словами, в которых оно встречается в языке. На этом факте как раз и строится обучение семантических агентов компании НейрОК.

Основная проблема здесь состоит в том, что при этом мы должны, подобно Мюнхаузену, «приподнять себя за волосы» (английский вариант – за шнурки ботинок, откуда и происходит название методики «bootstrap»). Действительно, значение каждого слова определяется его контекстом, состоящим из других слов, значения которых, в свою очередь, определяются их собственным контекстом, включающим и значение данного слова.

порядков. Например, амбициозный проект Netscape Open Directory Project [3] создания тематического дерева Интернет «всем миром», в данный момент насчитывает порядка 300 000 тематических рубрик.

Рис. 3. Bootstrap-алгоритм формирования самосогласованного набора семантических категорий при обучении на коллекции документов.

Поэтому алгоритм обучения построен на циклической схеме постепенного приближения к самосогласованной системе разложения всех слов по набору базовых категорий, отражающему статистику их совместного употребления в обучающей выборке (Рис. 3).

Зададимся неким фиксированным числом базовых категорий K, в зависимости от желаемой степени подробности семантического описания текстов. Семантику употребления каждого слова в языке будем кодировать разложением этого слова по категориям. Вначале это разложение полагается случайным.

После этого мы можем переопределить семантические векторы всех слов, в зависимости от частоты их употребления в том или ином кластере, и начать новый цикл погружения и кластеризации документов в новом, подправленном семантическом пространстве. Если на начальном этапе (случайное кодирование) все слова были в среднем равноудалены друг от друга, то уже после первой итерации кодирование слов будет в какой-то степени отражать их смысловую близость: слова, характерные для одних и тех же кластеров будут кодироваться сходным образом. Каждая следующая кластеризация будет более точной, т.к. будет учитывать смысловую близость слов, выявленную на предыдущих итерациях.

Можно показать, что, при соответствующем определении метрики, этот итерационный процесс сходится к локальному максимуму взаимной энтропии лексического, $P(w|d)$, и семантического, $P(k|d)$, представления документов.[3] То есть, сжатие информации при переходе из высокоразмерного в более низкоразмерное векторное пространство происходит с минимальными потерями. Соответственно, такое семантическое представление слов максимизирует количество информации о характере их употребления в языке.

Эти знания о структуре языка, выявленные в процессе обучения, отражают статистику совместного употребления слов. Так, поскольку различные словоформы одного и того же слова обычно употребляются в одном и том же тематическом контексте, их семантические векторы будут практически совпадать. Поэтому при семантическом поиске мы можем не заботиться о том, в какой именно форме употребляется то или иное слово в искомом документе. То же относится и к синонимам: неважно какими именно словами выражена та или иная мысль – ее семантическое представление практически не зависит от конкретного выбора терминов (**Рис. 4**).

[3] Где $P(w|d)$ – спектр терминов документа, а $P(k|d)$ – разложение документа по семантическим категориям, трактуемые как распределения вероятностей. Соответственно, взаимная энтропия двух распределений вероятности $I(W,K)=H(W)+H(K)-H(W,K)$ определяет информацию об одном из распределений, содержащейся в другом из них.

Рис. 4. Семантические векторы слов отражают смысловые связи между ними, выявленные в обучающей выборке документов.

В итоге, семантические фильтры оказываются гораздо «тоньше» лексических. В последнем случае для тонкой настройки на тему мы должны подобрать достаточно обширный набор терминов, который присущ данной тематике и наилучшим образом отличает ее от остальных. Этот набор должен использовать (i) заблаговременно составленный *тезаурус* языка, концентрирующий знания экспертов-лингвистов о словоформах и синонимах, (ii) знания экспертов в данной предметной области, ассоциирую-

щие употребляемые в ней термины.[4] Семантическое обучение автоматизирует получение этих знаний, позволяя обходиться без дорогостоящих экспертов.

Обучение семантике любого языка, в любой предметной области занимает даже на коллекциях из миллионов документов всего лишь несколько часов времени обычного персонального компьютера, т.е. обходится практически даром.

Описанный выше алгоритм обучения можно обобщить на выявление в структуре языка иерархической системы семантических категорий. В итоге, любая коллекция документов, независимо от ее языка и тематики, может быть автоматически структурирована – разбита на тематические директории, аннотированные наиболее значимыми словами. Соответственно, и потоки входящих документов будут автоматически раскладываться по этим тематическим директориям.

Semantic Explorer

Проиллюстрируем возможности семантического анализа данных на примере продукта фирмы НейрОК **Semantic Explorer 2**, предназначенного для автоматической кластеризации документов, рубрикации новых поступлений и поиска информации в больших хранилищах текстовой информации [4].

На стадии обучения сервер системы **Semantic Explorer 2** производит настройку своего семантического блока на «диалект» данного хранилища, выявляя базовые семантические категории. Автоматически формируется аннотированное дерево тематических категорий, по которым раскладываются все имеющиеся в хранилище документы.

К примеру, на **Рис.5** показано слева отображение тематического дерева на клиентской части комплекса **Semantic Explorer 2**, а справа – в панели результатов – содержимое одного из кластеров.

[4] Например, что А.Чубайс связан с РАО ЕЭС, а Р.Вяхирев – с РАО Газпром.

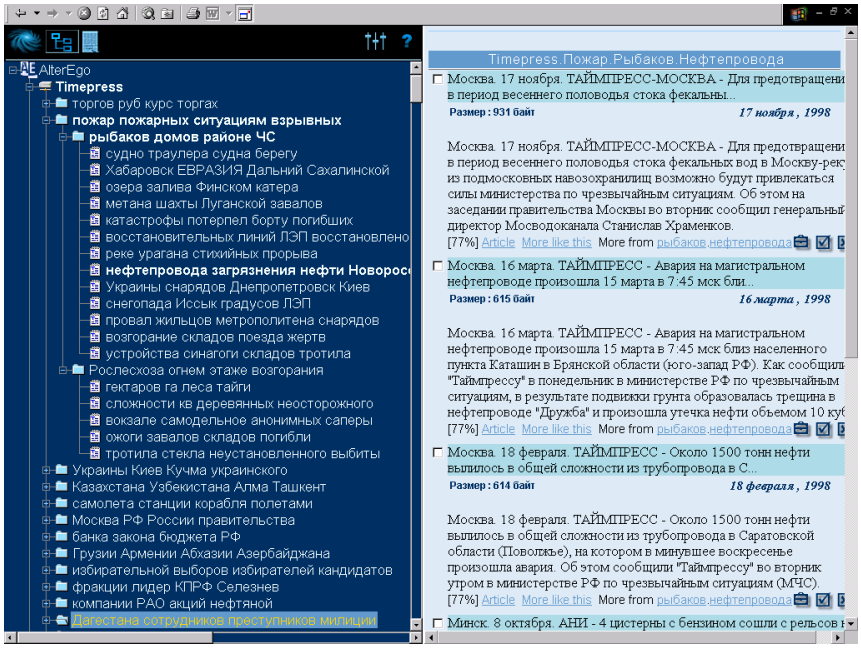

Рис.5. Иерархическая кластеризация выборки из 140 тысяч сообщений российских информационных агентств в системе Semantic Explorer 2. Слева – полученная в результате самообучения система тематических категорий. Справа – документы из выбранного (подсвеченного) кластера.

Поскольку все документы кодируются точками в семантическом пространстве, для каждого документа очень легко найти документы, ближайшие к нему по содержанию. Использование документов в качестве поисковых фильтров позволяет производить весьма тонкую подстройку поискового запроса.

Причем, найденные таким образом документы могут находиться и в других тематических категориях, поскольку ни одна классификация не в состоянии отразить истинного расположения документов в многомерном семантическом пространстве. Например, на **Рис. 6** документ про аварию на

магистральном трубопроводе принес документы из смежных кластеров, посвященных восстановительным работам и стихийным бедствиям.

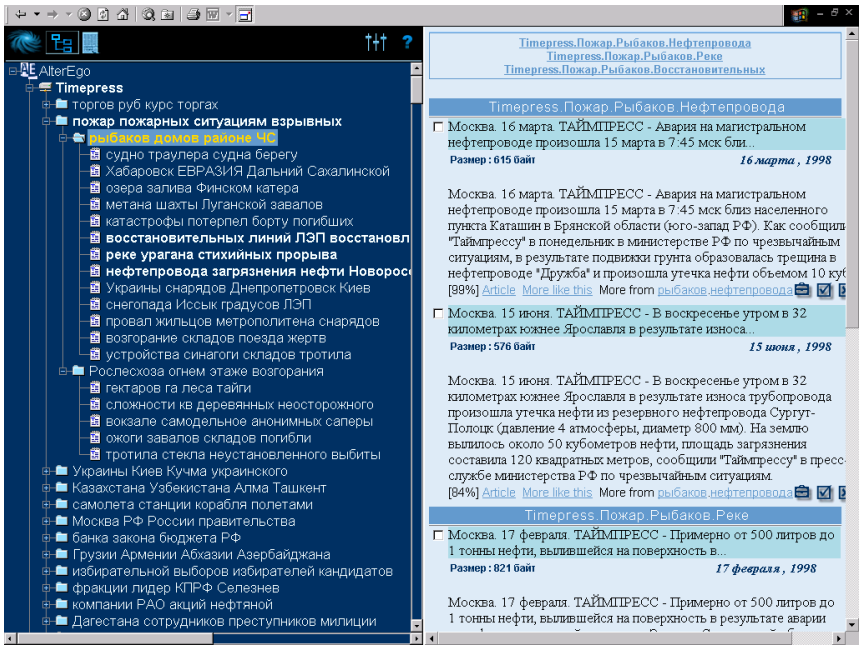

Рис. 6. Поиск похожих документов приносит документы из смежных кластеров.

В системе **Semantic Explorer 2** кроме семантического есть еще и лексический поиск. Семантическое дерево позволяет легко различать в каком именно контексте употребляется тот или иной термин в различных частях базы данных. Например, поиск по ключевому слову «Чубайс» приносит документы и из кластера «Москва РФ России», и из кластера «Компании РАО акции», отражающие соответственно деятельность Анатолия Борисовича как политического деятеля и бизнесмена (**Рис. 7**).

Рис. 7. Ключевые слова встречаются в разных контекстах.

Кроме привычного «одномерного» отображения дерева тематических каталогов, в **Semantic Explorer 2** реализованы и двумерные карты, дающие наглядное представление об объеме и относительном расположении различных тематических кластеров. Навигация по коллекции в этом случае напоминает полеты в виртуальных галактиках документов (**Рис.8**).

Установив на входе **Semantic Explorer 2** робота, выкачивающего и разбирающего содержимое http-страниц, мы получаем средство автоматической каталогизации Интернет – автоматический Yahoo! Причем, различные языковые сегменты Интернет будут автоматически формировать разные ветви этого дерева. И все это – без привлечения экспертов и без априорного знания многочисленных языков.

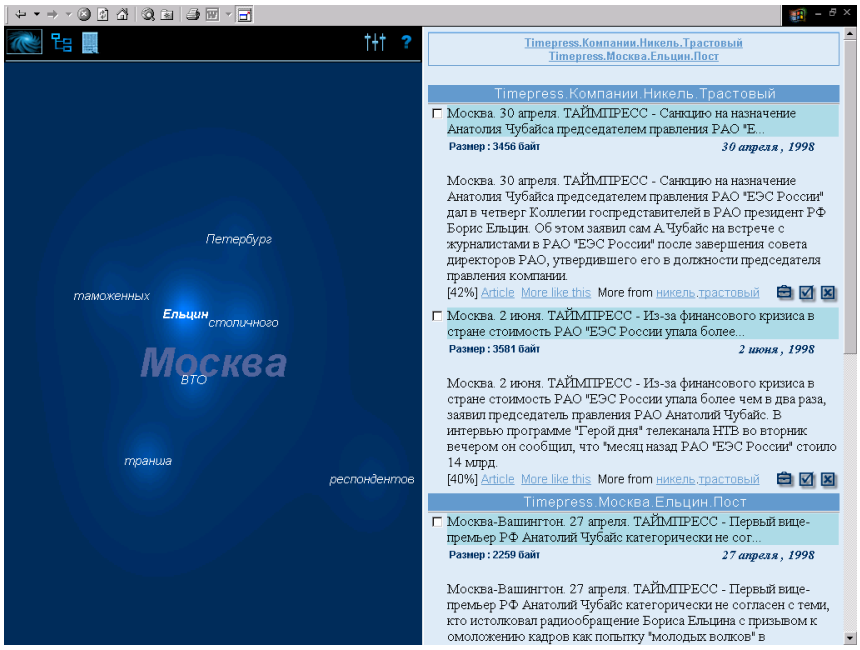

Рис.8. Галактики документов в системе Semantic Explorer 2.

На самом деле, конечно, это типичная утопия. Как говорил Козьма Прутков: «Нельзя объять необъятное». Мы опять возвращаемся к основной теме данной статьи – распределенным поисковым системам. Содержание быстро растущей и постоянно изменяющейся глобальной Сети способна проиндексировать только такая же растущая вместе с ней сеть постоянно обучающихся поисковых агентов, каждый из которых знает лишь частицу всеобщей Семантики.

Как может быть устроена такую сеть – тема следующего раздела.

Сеть семантических агентов

Представим себе, что на персональном компьютере каждого пользователя Сети живет его поисковый агент, обучающийся семантике по мере выполнения запросов своего хозяина. Поскольку человек знает обычно не так много языков и имеет ограниченный набор интересов, семантическое пространство такого персонального агента будет относительно небольшим. В нем будут представлены лишь те семантические категории, которые нужны данному пользователю.

Это семантическое пространство служит для индексирования найденных документов, заинтересовавших или могущих заинтересовать пользователя, а также – и это самое важное – *источников* получения таких документов – других агентов (Рис. 9).

Рис. 9. Семантическое пространство поискового агента используется для индексации документов и их источников – других агентов.

К ним он обращается, если в локальном индексе нужных документов не оказалось. Причем, семантические координаты ссылок на других агентов отражают тематику, по которой к нему лучше обращаться. В свою очередь, другие агенты могут обратиться к данному в поисках информации по его «профилю».

Все вместе агенты образуют тесно переплетенную связями поисковую сеть, охватывающую в совокупности все присутствующие в Интернете тематики на всех используемых в нем языках. Напомним, что совокупное дисковое пространство всех ПК составляет б*о*льшую часть всей мировой

дисковой памяти. Так что потенциал такой поисковой сети по определению сопоставим с объемом информации всего глубинного Интернета.

Подобная сеть поисковых агентов является идеальным средством направленного распространения информации в Сети. Агенты получают информацию только по четко определенной текущим интересом пользователя тематике. Даже реклама, распространяемая в такой сети наряду с документами, будет всегда четко коррелировать с тематикой запроса. Впрочем, легко доступная по такой сети коммерческая информация может существенно снизить роль обычной рекламы.[5]

Свободное компьютерное время агент использует для наращивания своих знаний – выкачивает новости, ищет новых агентов, улучшает структуру своих связей, чтобы при необходимости обращаться в первую очередь к тем агентам, которые принесут более качественную информацию.

Чтобы цепочка запросов была относительно короткой, а поиск, соответственно, быстрым, необходимо обеспечить достаточную плотность межагентских связей. Типичное кратчайшее число звеньев между произвольными точками сети называется ее радиусом. По существующим оценкам, радиус гипертекстовой среды WWW превышает 20 – слишком много, как мы знаем из личного опыта, для того, чтобы следуя гиперссылкам найти нужный документ. Для сравнения, радиус сети человеческих знакомств равен приблизительно 6. Т.е. каждый человек знаком с каждым другим в среднем через цепочку из 6 личных знакомств. Этот эффект известен под названием «малого мира». Радиус сети R легко оценить: $N \sim \kappa^R$, где N – объем, а κ – типичное число связей между ее звеньями. Если в качестве N принять число тематических кластеров-«нейронов» в агентской сети [6], равное число агентов A помноженное на число нейронов в «мозге» типичного агента K, $N = KA$, то связность такой семантической нейронной сети оценивается как:

[5] «Реклама для роботов» уже широко используется в Сети для «накручивания» индекса популярности сайтов. Однако, она может играть и конструктивную роль.

[6] Т.к. мы хотим получать быстрый ответ на вопрос по любой тематике.

$$\kappa \sim (KA)^{1/R}$$

Если ориентироваться на мир людей, в котором $R \sim 6$, то при реалистических оценках $K \sim 10^2$, $A \sim 10^8$, получаем $\kappa \sim 50$. Это значит, что типичный агент должен иметь около $\kappa K \sim 5000$ ссылок на других агентов. Естественно, что механизм установления всех этих связей должен быть полностью автоматизирован.

На этом, центральном, моменте стоит остановиться подробнее. Как сделать так, чтобы запрос не «утонул» в сети поисковых агентов, а всегда приносил полезную информацию? Как улучшать топологию поисковой сети в целом в результате локального обучения каждого отдельно взятого агента?

Экономика семантических агентов

Достижение желаемого глобального поведения сети при локальном характере функционирования и обучения каждого нейрона этой сети – краеугольный принцип нейрокомпьютинга. Каждый нейрон получает входную информацию для переработки от одних нейронов и передает свои результаты по сети другим нейронам. Сигнал об успешности своей деятельности он получает от последних и передает далее по сети в обратном направлении тем, от кого получает входные данные. Причем, ошибка распределяется по сети не равномерно, а пропорционально вкладу каждого нейрона в ее происхождение. Такой метод обучения – метод обратного распространения ошибки (error backpropagation) – позволяет организовать градиентный поиск глобального оптимума в распределенных системах, где каждый обладает лишь частицей информации получаемой от своего ближайшего окружения.

Так устроены нейронные сети, но так же функционирует и экономика! Каждый экономический субъект потребляет услуги одних и производит услуги другим субъектам экономики – будь это на уровне стран, фирм или отдельных индивидуумов. Сигналом, характеризующим вклад этого субъекта в общее дело, являются деньги – отсутствие таковых свидетельствует об ошибочном экономическом поведении. Денежные потоки от

конечных потребителей вниз по производственной цепочке вполне аналогичны обратному распространению ошибки. Причем, как и там, потенциальные выигрыши или потери экономических субъектов пропорциональны их вкладу в конечный результат. Обучение, основанное на таком простом и понятном каждому локальном сигнале, как его собственный доход, оказывается, как мы знаем, достаточно эффективным, для обеспечения жизнедеятельности всего человеческого общества.[7]

Успешность экономического регулирования наводит на мысль, что подобным же образом можно организовать и сеть поисковых агентов. Услуги, которые агенты оказывают друг другу, можно измерять единым денежным эквивалентом. Пусть эта валюта называется, например, *нейро*.

Базовая поисковая услуга сводится к поставке одной ссылки на документ или агента по заданной тематике с заданной точностью. Последнее означает, что семантическое расстояние между запросом и выданной по запросу ссылкой не должно превышать некоторого порога. Запрос может иметь также ограничения на тип и возраст документа. Таким образом, последние новости будут отделяться от архивных документов. Стоимость такой услуги определяется в процессе взаимодействия поисковых агентов, в котором каждый из них старается максимизировать собственную выгоду. В результате возникает некая равновесная взаимовыгодная цена, которая является экономическим обоснованием данной связи.

[7] Хотя каждый человек, бесспорно, является гораздо более сложной системой, чем простой формальный нейрон.

Рис. 10. Схема функционирования и обучения поисковой сети.

Попробуем сформулировать основные положения «экономики агентов» и посмотрим как может функционировать подобная поисковая экономика:

- Каждый поисковый запрос должен сопровождаться соответствующим денежным обеспечением. Предполагается, что есть минимальный денежный квант, скажем 1 *нейро*. Это обеспечит затухание любого запроса в сети, предотвращая его бесконечную репликацию.
- Целью агента является максимизация своего *дохода*, т.е. притока денег. Будем полагать, что полученные деньги он не ко-

пит, а немедленно реинвестирует в улучшение своего состояния с таким расчетом, чтобы максимизировать приток денег в будущем.

- Состояние агента характеризуется имеющимися в его локальном индексе семантическими ссылками на документы и других агентов, разложенными по базовым категориям-нейронам, причем как нейроны, так и ссылки имеют свой адаптивный рейтинг – оценку ее потенциальной доходности. Изменением этих рейтингов, а также путем получения новых ссылок агент старается улучшить свое экономическое поведение.

Допустим, к агенту поступил запрос на один документ по заданной тематике с некоторым денежным обеспечением. Что должен делать агент с точки зрения экономической целесообразности?

- Он должен использовать полученную информацию для коррекции своих оценок платежеспособного спроса на данную тематику. Это достигается коррекцией рейтингов соответствующих нейронов агентского мозга (индекса). Эта информация будет позже использована им для пополнения индекса по наиболее прибыльным тематикам.
- Он должен постараться удовлетворить поступивший запрос чтобы не уменьшить свой рейтинг у заказчика вопроса, что привело бы к уменьшению ожидаемой прибыли. В первую очередь следует использовать ссылки, имеющиеся в его собственном индексе. В этом случае прибыль можно будет позже вложить для оптимизации своего состояния.
- В случае, если агент не в состоянии удовлетворить запрос за счет своего индекса[8], он пытается использовать потенциал своих связей с другими агентами, приносившими ему ранее документы по данной тематике, переадресовав запрос одному (или нескольким – в зависимости от имеющихся средств) из

[8] Напомним, что агент всегда имеет формальный критерий требуемой релевантности ответа, поступающий вместе с запросом.

них с наивысшим рейтингом. Тем самым, он повышает вероятность успешной отработки запроса и поддержания своего реноме у заказчика.

- По результатам полученных ответов агент подправляет рейтинги тех нейронов, к которым он обращался, и выдает, если сможет, ответ своему заказчику.
- Переадресовывая запрос, агент может оставить часть средств себе в виде «комиссионных». Полученную за счет этого, а также за счет использования своего локального индекса прибыль агент использует для получения новых ссылок на документы и агентов по наиболее прибыльным, по его текущим оценкам, тематикам.

Таким образом, будут автоматически укрепляться связи между теми агентами, которые чаще приносят правильные ответы на наиболее часто обращаемые к ним запросы.

Источником поисковой валюты являются конечные заказчики – хозяева своих агентов. На них, как правило, и будут ориентироваться агенты при оптимизации своих связей. Однако, возможны и другие варианты, при которых агенты начнут спонтанно специализироваться на какой-то популярной в сети тематике, ориентируясь, в конце концов, на интересы (и деньги) других пользователей.[9]

Здесь мы подходим к самому деликатному вопросу – кто будет ответственен за эмиссию циркулирующей в сети поисковой валюты? Если дать эту функцию «на откуп» хозяевам агентов, то последние будут стараться нивелировать влияние других агентов на своего, например, бесконтрольно увеличивая плату за свои запросы. Тем самым, общественно-полезная функция таких агентов будет подавляться. Множественность эмиссионных центров приведет к разбалансированию системы и сведет на нет все ее преимущества.

Выходов из этой ситуации, по большому счету, два. Первый, «административный», – лимитировать количество *нейро* на поисковый запрос на

[9] С наибольшей вероятностью такую популярность обретут агенты с самого начала ориентированные своим хозяином на соответствующую тематику.

уровне программного кода. Однако, это снизит возможности пользователя акцентировать свои запросы, например, если ему нужен целый обзор по определенному вопросу. Альтернативный, «рыночный», подход – превратить поисковую валюту в настоящую, введя плавающий обменный курс, определяемый глобальным спросом на поисковые услуги. В этом случае пользователь будет сам распоряжаться своими деньгами, выделенными на подключение его к агентской поисковой сети. Курс *нейро* при таком подходе будет реальным мерилом того, насколько такая агентская сеть окажется полезной для жизни пользователей в целом.

В этом последнем случае хозяева популярных агентов, берущих на себя общественно-полезные функции, будут иметь реальную компенсацию за это в виде зарабатываемых их агентами реальных денег. Более того, не исключено, что появятся агенты, приносящие своим хозяевам заработок, достаточный для того, чтобы посвящать все свое время администрированию подобных агентов.

Агенты и бизнес

Выше мы описали экономику поисковой сети в предположении, что пользователи будут сами оплачивать услуги по поиску информации в Сети. Между тем, все уже давно привыкли, что подобные услуги в Интернете бесплатны. На самом деле, конечно, кто-то все равно оплачивает издержки по «бесплатным» службам Интернет. Этот «кто-то» – коллективный рекламодатель и стоящий за его спиной большой бизнес.

Сегодня эта реклама оседает в Интернет-порталах, аккумулирующих на себе внимание миллионов пользователей, за которое им и платят рекламодатели. Однако, агентская сеть сможет распространять рекламу гораздо более направленно и эффективно, чем вещающие в 4π порталы.

Рис. 11. Финансирование поисковой сети за счет рекламы.

Представим себе, что все рекламные деньги, вместо того, чтобы оседать на порталах, будут распределяться среди распространяющих ее агентов в поисковой сети. За каждый доставленный пользователю баннер или купон скидки в электронном магазине агент-посредник получает некоторое количество *нейро*. Общее количество рекламы в Интернет в 2004 году оценивается в $30 миллиардов [5]. Эти деньги могут обеспечить вполне сносное существование ($30 тысяч в год) примерно миллиону активных общественно-полезных агентов, на поддержание которых имеет смысл тратиться их хозяевам. Такое «поисковое ядро» из миллиона тематических поисковых серверов намного превосходит по своим ресурсам суммарную мощность нынешних универсальных поисковиков.

Этот количественный скачок в производительности поиска сможет обеспечить качественно иной уровень информационного обеспечения в Интернет.

Новые возможности

Действительно, даже если принять, что суммарная производительность распределенного поиска всего лишь в 10^3 превышает централизованные, это уже означает, что распределенный индекс будет в тысячу раз актуальнее последних. То есть вместо недельных или даже месячных лагов индексирования, характерных для лучших из существующих поисковых серверов, мы получим практически мгновенную индексацию в реальном времени. Любой документ, помеченный пользователем «для общего пользования» будет в тот же момент доступен всем агентам из ближайшего сетевого окружения. Т.е. тем, которые обычно обращаются к нему за документами с похожим содержанием, и, следовательно, кому он будет, скорее всего, интересен.

Таким образом, кардинально упрощаются публикации «для единомышленников». Сейчас роль концентраторов подобных публикаций играют специализированные журналы. Но они, во-первых, покрывают достаточно большие тематические области, и во-вторых, разные авторы предпочитают публиковаться в разных журналах. Агенты позволят автоматически объединять профессиональные сообщества с очень тонкой настройкой по интересам. И не только профессиональные. Всевозможные увлечения и хобби, локальные сообщества, объединяющие тех, кто хочет участвовать в местной общественной жизни или интересуется успехами местных спортивных команд.

В общем, агентская сеть предоставляет тот недостающий сегодня инструмент общения «человек-сообщество», о котором мы упоминали в начале этой лекции. В пирамиду человеческих отношений, на вершине которой – однородное «вещание» на всех средствами масс-медиа, а в основании – личное общение с близкими по жизни и работе людьми, органично встраивается серединное звено – общение с единомышленниками (Рис. 12).

Интенсивное общение между агентами, обслуживающими информационные интересы своих пользователей, в результате «закулисного» самоорганизующегося процесса, прозрачно для пользователя помещает его в соответствующее его интересам место в глобальной информационной сети. У человека автоматически формируются некие «социальные координаты», в которых он имеет ближнее окружение в лице хозяев агентов, интересующихся сходными вопросами. В зависимости от своих наклон-

ностей, он может проигнорировать или воспользоваться этими новыми возможностями установления социальных контактов и общения с единомышленниками.

Рис. 12. Агенты автоматически формируют среду для эффективного общения по интересам.

Простота публикации своих мнений, которые наверняка прочтут, и на которые, возможно, отреагируют люди из «ближнего круга» – новое качество, имманентно присущее самообучающейся агентской среде. Можно ожидать, что количество такого рода полемических публикаций, вопросов и ответов на них существенно возрастет по сравнению с тем, что мы сегодня имеем в USENET. Агенты способны обеспечить гораздо более естественный интерфейс общения по интересам, чем USENET. А удобный интерфейс, как показывают примеры Windows и графических браузеров, является непременным условием массовой компьютерной технологии.

Обсуждение и заключение

Итак, подводя итоги. Мы начали нашу лекцию с констатации того, что сегодняшний Интернет находится в стадии становления и еще далек от совершенства. Глобальные поисковые серверы индексируют лишь поверхностный слой доступной информации, нет справочно-поисковой службы, объединяющей многочисленные доступные из Сети базы данных. Недостаточно развиты средства получения персональных новостей, формирования сообществ по интересам и общения в таких сообществах. Общение Бизнеса с потенциальными покупателями, в частности, распространение рекламы, использует заимствованную у телевидения модель вещания с больших порталов, привлекающих к себе массовую аудиторию. Иными словами, принципиально новые возможности Интернет как программируемой среды общения и распространения информации используются пока далеко не полностью.

Описанная в лекции концепция создания сети персональных агентов призвана хотя бы частично улучшить существующую ситуацию. Эта идея не оригинальна – в последнее время она буквально носится в воздухе. Достаточно упомянуть несколько недавних крупных peer-to-peer[10] проектов, идеи которых перекликаются с нашими: Napster [6], Freenet [7], Gnutella [8], MojoNation [9], OpenCOLA [10].

Что отличает от перечисленных данный проект, реализуемый российской фирмой НейрОК [4], так это существенная опора на технологии машинного обучения. Во всех предлагаемых до сих пор системах агенты не распознают семантику контента и связываются друг с другом безотносительно содержания индексируемых ими документов. В результате, поиск в таких сетях требует перебора чуть ли не всего распределенного индекса, что, естественно, неприемлемо. По нашему мнению только самообучающиеся семантические агенты путем самоорганизации способны выстроить достойную Интернета постоянно адаптирующуюся семантическую сеть поиска и распространения информации.

[10] Peer-to-peer – равный с равным. Архитектура сети без центрального управления потоками данных.

Литература

[1] http://www.sims.berkeley.edu/how-much-info/internet.html

[2] http://www.sims.berkeley.edu/how-much-info/charts/charts.html

[3] http://dmoz.org/about.html

[4] http://www.neurok.ru

[5] http://www.geocities.com/CollegePark/Quad/5905/WordFile.html

[6] http://www.napster.com

[7] http://freenet.sourceforge.net/

[8] http://gnutella.wego.com/

[9] http://www.mojonation.com/

[10] http://www.opencola.com/

Дополнительные источники информации

Интернет

"Sizing the Internet," *Cyveillance*:
http://www.cyveillance.com/resources/library.asp

"The Deep Web: Surfacing Hidden Value," *BrightPlanet LLC*:
http://www.completeplanet.com/Tutorials/DeepWeb/index.asp

"Web Surpasses One Billion Documents," *Inktomi Corp.*:
http://www.inktomi.com/new/press/billion.html

"Size of the Web: A Dynamic Essay for a Dynamic Medium," *The Censorware Project*:
http://censorware.org/web_size/

"State of the Internet 2000," *United States Internet Council & ITTA Inc.*:
http://usic.wslogic.com/intro.html

"Domain Statistics," *DomainStats.com*: **http://www.domainstats.com**

"Email Facts," *24/7 Media*:
http://www.247media.com/research/trends/email.html

"Like It Or Not, You've Got Mail," *BusinessWeek*:
http://businessweek.com/1999/99_40/b3649026.htm

"Year-End 1999 Mailbox Report," *Messaging Online*, **http://www.messagingonline.com/**

"Top 1000 Usenet Sites" *Freenix*, **http://www.freenix.org/reseau/top1000/**

Программные агенты

Наиболее полная подборка материалов о программных роботах в Интернет**:** **http://www.botspot.com/**

Удобный мета-поисковый агент: **http://www.copernic.com/**

Персональный ассистент Adalyn: **http://www.adalyn.com/**

Коммерческие агенты компании Artificial Life, Inc.: **http://www.artificiallife.com/**

Блуждающие агенты компании Tryllian: **http://www.tryllian.com/**

Список shopping агентов:
http://dir.yahoo.com/Business_and_Economy/Shopping_and_Services/Retailers/ Virtual_ Malls/Shopping_Agents/

Peer-to-peer

Организация распределенных вычислений в Интернет, *Global Grid Forum*:
http://www.gridforum.org/

Распределенная сеть индексирования локальных файлов, *iMesh*: **http://www.imesh.com/**

Разработка стандартов P2P для Интернет, *Peer-to-Peer Working Group*,
http://www.peer-to-peerwg.org/

Мониторинг новостей по P2P: **http://www.peerprofits.com/**

Статья о «врожденных» недостатках Gnutella, *"The Gnutella paradox"*:
http://salon.com/tech/feature/2000/09/29/gnutella_paradox/

Статья создателя *Lotus Notes* о P2P: **http://www.zdnet.ru/news.asp?ID=1920**

Сергей Александрович ШУМСКИЙ, кандидат физико-математических наук, старший научный сотрудник ФИАН им. П.Н.Лебедева РАН, заместитель генерального директора ООО «НейрОК». Научные интересы – физика плазмы и термоядерного синтеза, статистическая механика распределенных вычислений, теория и приложения нейрокомпьютинга. Автор более 40 научных работ и одной монографии.