РОССИЙСКАЯ АКАДЕМИЯ НАУК

МИНИСТЕРСТВО ПРОМЫШЛЕННОСТИ, НАУКИ И ТЕХНОЛОГИЙ РОССИЙСКОЙ ФЕДЕРАЦИИ

ВСЕРОССИЙСКИЙ ИНСТИТУТ НАУЧНОЙ И ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ (ВИНИТИ)

НАУЧНО • ТЕХНИЧЕСКАЯ ИНФОРМАЦИЯ

Серия 2. ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ И СИСТЕМЫ

ЕЖЕМЕСЯЧНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ СБОРНИК

Издается с 1961 г. № 1 Москва 2001

ОБЩИЙ РАЗДЕЛ

УДК 002.513.5+004.832

С. В. Попов

Поиск информации и принятие решений

Рассматриваются противоречия между различными направлениями теоретической информатики и отсутствие единого фундамента, на котором должно строиться информационное общество. В качестве основного нерешенного вопроса выделяется определение критериев эффективности информационных систем. Предлагается модель построения информационной системы, в которой критерием оптимизации функционирования является оценка вероятности изменения точки зрения пользователя на стоящую перед ним проблему в результате получения новой информации.

I. ИНФОРМАЦИОННОЕ ОБЩЕСТВО — КРИЗИС НЕЭФФЕКТИВНОСТИ

В докладе на II Всероссийской научной конференции "Россия — XXI век" под названием "Технократический подход к информатизации общества — источник угроз национальной безопасности России" проф. В. Е. Лепский определил те признаки нарождающегося информационного общества, которые являются, по его мнению, потенциальными источниками угроз тотального манипулирования человеком и сообществами:

- анонимность источников информации;
- манипулирование навигацией пользователей;
- манипулирование сообществами пользователей;
- манипулирование предпочтениями пользователей и их потребностями;
- стрессовые воздействия и возникновение синдрома зависимости от киберпространства;
- сочетание безграничных информационных просторов и примитивных средств навигации, монополизированных узкой группой фирм.

По мнению проф. Лепского, "доминанта технократического подхода к информатизации общества завела эти процессы в глубокий кризис".

На мой взгляд, дело обстоит еще хуже: даже с точки зрения технократического подхода, фундамент, на котором строится так называемое информационное общество, а вернее, отсутствие этого фундамента может привести этого фантастически быстро развивающегося монстра к "коллапсу".

Сегодняшнее состояние дел в информатике напоминает начальную стадию развития и распространения паровых машин в конце XVII — начале XVIII вв., когда еще не осознав понятий "теплота", "энергия", "работа", "энтропия", "к.п.д.", тогдашнее общество смело и решительно внедряло это изо-

бретательское новшество в практику. Прошло, по крайней мере, еще сто лет, пока, благодаря трудам Карно, Клаузиуса, Больцмана и др., не появилась наука "термодинамика", позволившая осмысленно повысить эффективность тепловых двигателей, оптимизировать параметры их работы и научиться оценивать последствия их внедрения.

В настоящее время отсутствуют общепринятые теоретические основы информатики:

1. Теоретические работы Винера, Шеннона, Эшби, Колмогорова, Котельникова, Бриллюэна и др. не получили законченного развития в конце XX в. и не превратились в стройную теорию.

2. Экспериментальные открытия библиотекарей-энтузиастов, так называемые "законы" Ципфа, Лотки, Брэдфорда, так и остались теоретически неосмысленными. Знаменитый Крэнфильдский проект, направленный на тотальный сбор статистических данных, касающихся оценки эффективности информационных систем, привел к противоречивым результатам и больше не повторялся, в первую очередь, в силу своей дороговизны.

3. Не только не ослабилось, но и усилилось взаимонепонимание различных "теоретических школ". В первую очередь это касается кибернетического (винеровского) направления, с одной стороны, и библиометрически-лингвистического — с другой.

Представители каждого из этих направлений стали говорить о выдающихся приложениях своих теорий и о том, что вот-вот будет разработана "идеальная информационная машина".

Если винеровская школа грозит в ближайшие годы создать "искусственный интеллект", способный заменить человека при решении не только тривиальных, но и творческих задач, то лингвистическая школа, проникшая гораздо глубже в основу всякого мышления, — язык, пессимистически взирая на эти механистические попытки, стала "заглядываться" на "мистическую информатику" — бескомпьютерный обмен информацией, информационные поля, ритмодинамику и т. д. При этом представители обеих школ даже не удосуживаются обратить друг на друга внимание и практически не спорят. В стороне от этих двух столбовых дорог оказалось "негэнтропийное направление", парадоксально превратившееся из "информационной физики" в узаконенный раздел математической теории вероятностей, несмотря на то, что Клод Шеннон достиг выдающихся результатов именно в сугубо практическом применении своей теории и до конца жизни считал себя простым инженером. То же самое произошло с теорией статистических решений, которая вообще незнакома большинству создателей "электронных библиотек", занимающихся оптимизацией информационного поиска и вынужденных постоянно "изобретать велосипеды" типа ошибок первого и второго рода, давно изученных в рамках этой теории, и называть их другими именами — "полнота", "специфичность", "точность" и т. д.

При всем этом общественное мнение "подогрето" ожидаемыми успехами "информационного общества", поголовной компьютеризацией, Internet, DVD, фантастически емкими носителями, мультимедиа, виртуальной реальностью и т. д.

Апологеты этого общества, скорее всего, умышленно стараются не афишировать тот факт, что все перечисленные достижения — следствие развития все той же физики (электроники, приборостроения, связи и т. д.), а совсем не фундаментальной информатики, которой на сегодняшний день просто нет.

Кое-кому приходит мысль, что она вообще не нужна. Как это ни парадоксально, многим владельцам информационных систем не выгодна их эффективная работа, так как, по придуманным кем-то правилам, пользователь платит не за качество поиска, а за время, проведенное в системе, и/или за количество найденных документов, независимо от того, полезны они ему или нет. Представьте картину, когда бы первые "паровые машины" стали одновременно неэффективно работать во всех домах граждан Америки, Азии, Европы и даже Антарктиды... Думаем, через год не было бы ни граждан, ни домов, ни самих этих континентов. Что-то похожее ожидает нас в ближайшее время в духовной сфере. Причем "бомба" будет нейтронной, т. е. останутся континенты, дома, не будет только нас с вами или мы станем совсем другими, сидя у своих "паровых" компьютеров.

Великие физики прошлого, разработав такие глубокие понятия, как "энергия", "энтропия", "работа", и установив количественные законы, описывающие их взаимосвязь, стремились ответить всего лишь на два простых, но очень важных вопроса, касающихся технической эффективности тепловых машин:

1) можно ли со стопроцентной эффективностью превратить тепловую энергию в работу?

2) если нет, то при каких условиях коэффициент полезного действия тепловой машины будет максимальным и чему он будет равен?

Аналогичные вопросы стоят сейчас перед разработчиками информационных систем. Платон сказал: "Если ты знаешь то, что хочешь найти, то зачем тебе это искать, — ты ведь и так это знаешь, а если ты не знаешь то, что хочешь найти, то как ты это найдешь?" Этот парадокс остается не разрешимым до сих пор. Успехи новых информационных технологий сегодня определяются только возможностью быстро получить фактографическую информацию при условии, что возможно полностью формализовать свой информационный запрос: расписание поездов, наличие лекарств в аптеках, цены на продукцию и т. д.

Однако основная информационная проблема, стоящая перед каждым из нас каждый день, связана с принятием решений в условиях неопределенности. Способствуют ли современные информационные системы повышению эффективности принимаемых решений? Думаю, что нет. Одно дело, когда ты точно знаешь название лекарства и просто выбираешь ближайшую аптеку, в которой оно есть, и совсем другое дело, когда нужно поставить диагноз и подобрать подходящее лекарство из бесконечного моря существующих препаратов. Неполнота полученных из информационной системы данных может в этом случае привести к катастрофическим для больного последствиям.

Итак, информационной науке нужно ответить на два вопроса:

1) можно ли со стопроцентной вероятностью превратить имеющуюся информацию в безошибочное решение?

2) если нет, то при каких условиях это решение будет наиболее вероятным, и какова эта вероятность?

Не ответив на эти два вопроса, современное общество рискует попасть под гнет совершенно нового тоталитаризма — тоталитаризма информационного, когда большинство решений спускается незаметным образом из какого-то "анонимного центра" и навязывается нам не с помощью силы, а вследствие отсутствия у нас достоверной и полной информации.

II. ИНФОРМАЦИОННАЯ ПОТРЕБНОСТЬ И ИНФОРМАЦИОННЫЙ ЗАПРОС. ПОПЫТКА РАЗРЕШЕНИЯ КОНФЛИКТА

Всем известны слова — запрос и потребность. В информатике особую роль играют термины — *информационный запрос* и *информационная потребность*, придуманы также соответствующие свойства искомой информации — "релевантность" и "пертинентность". Скорее всего, в информационном обществе всякий запрос и всякая потребность будут информационными, но от этого жить легче не станет. Ведь всякий запрос не адекватен потребности и чаще всего не отражает ее полностью, а иногда и вовсе ей противоречит. Более того, в информационном обществе всякая потребность может обрести конкретный и единственный смысл — приобретение новых знаний для принятия корректных решений. Скорее всего, по мере развития информатики критерии оптимизации функционирования информационных систем должны будут опираться на определение количества приобретаемой человеком новой информации, а не на оценку соответствия выданных документов тематике запроса. Так что же такое "новая информация" и как определить меру ее "новизны".

Каждая проблема связана с ее описанием, чаще всего неправильным. Если мы составим словарь этого описания (Словарь 1) и, сравнив его со словарем некого огромного информационного массива, представляющего все порожденные человеком тексты, попытаемся выразить все эти тексты на Словаре 1, т. е. использовать во всех текстах только слова этого небольшого словаря, то огромное большинство текстов для нас как бы исчезнет, другая большая часть текстов будет содержать только отдельные слова из Словаря 1, еще некоторая часть только пары слов, еще меньшая — тройки и т. д. Некоторое количество текстов будет содержать почти все слова из Словаря 1, скорее всего, именно эти тексты или их фрагменты будут описывать проблему точно так же неправильно или недостаточно правильно. К сожалению, для клиента информационной системы и в "исчезнувших" текстах, и в текстах "одиночках", "двойках" и других может содержаться информация, способная изменить его взгляд на существующую проблему. Однако вероятность этого изменения будет разной, в зависимости от того, будет ли полученный текст из класса "одиночек", "двоек" или "троек".

Если рассматривать информацию мирового массива безотносительно к существующей проблеме, то выражение этой информации на Словаре 1 можно назвать "потерями информации". В "потерянном классе" (Класс 0) "не осталось" ни одного текста, в классе "одиночек" (Класс 1) остались некие тексты, состоящие из одного слова. Однако, если посмотреть на этот класс более внимательно, то его структура будет гораздо сложнее, чем у Класса 0. Если в Классе 0 все тексты стали одинаковыми — "никакими", то в Классе 1 появляются группы одинаковых текстов — группа текстов, состоящих только из Термина 1 Словаря 1, группа текстов, состоящих только из Термина 2 Словаря 1 и т. д. Обозначим количество текстов из первой группы Класса 1 — n_{11}, второй — $n_{21} \ldots n_{к1}$, где к — количество групп в Классе 1.

Разнообразие состава Класса 1 можно оценить шенноновским выражением негэнтропии:

$$H_1 = -\sum_1^k p_{i1} \ln p_{i1},$$

где

$$p_{i1} = \frac{n_{i1}}{n_1},$$

а n_{i1} — количество текстов группы i Класса 1, n_1 — полное количество текстов Класса 1.

Негэнтропия Класса 0 будет

$$H_0 = -\sum_1^1 1 \ln 1 = 0;$$

негэнтропия произвольного Класса j будет

$$H_j = -\sum_1^{k_j} p_{ij} \ln p_{ij}, \qquad (1)$$

где k_j — полное количество групп в Классе j;

$$p_{ij} = \frac{n_{ij}}{n_j},$$

где n_{ij} — количество текстов в группе i Класса j, n_j — полное количество текстов Класса j.

Негэнтропию Класса текстов, содержащих все слова из Словаря 1,обозначим H_m, где m — полное количество слов в Словаре 1. Состояние Класса m аналогично состоянию Класса 0, так как все тексты этого Класса "становятся" на Словаре 1 одинаковыми. Отсюда

$$H_m = -\sum_1^1 1 \ln 1 = H_0 = 0.$$

В Институте промышленного развития (Информэлектро) проведены многочисленные эксперименты с большими массивами информации (имитирующими "мировой поток") и со словарями, отражающими конкретные проблемы. Эксперименты показали, что распределения величин H_j по классам для разных случаев хорошо аппроксимируются нормальным (гауссовым) распределением.

Эксперименты показали также, что существует способ оценки степени влияния текстов, выбранных из определенных групп, на изменение взглядов пользователей на стоящую проблему и ее описание. Приблизительно это изменение можно оценить увеличением и/или уменьшением "словаря проблемы", т. е. переходом Словаря 1 в некий Словарь 2.

Обозначим это изменение $\Delta S_{<t>}$. $\Delta S_{<t>}$ — количество новых терминов, появившихся в "словаре проблемы", плюс количество терминов, убранных

пользователем из "словаря проблемы" после прочтения текстов большого массива, содержащих некий набор терминов $< t >$ из Словаря 1.

Эксперименты показывают, что

$$\Delta S_{<t>} \sim \sum_{1}^{m} p_{<t>j} H_j, \qquad (2)$$

где

$$p_{<t>j} = \frac{n_{<t>j}}{n_{<t>}},$$

здесь $n_{<t>j}$ — количество текстов, содержащих набор терминов $< t >$ из Словаря 1 в Классе j, n_j — полное количество текстов в "мировом потоке", содержащих набор терминов $< t >$, $\sim$ — знак пропорциональности.

Используя правило (2), информационная система способна наилучшим образом изменять взгляд пользователя на интересующую его проблему. Работа такой системы действительно направлена на поддержку принятия наиболее объективных решений в условиях неопределенности.

Итак, отвечая на вопросы, поставленные в первой части настоящей статьи, можно сказать следующее:

- имеющуюся информацию невозможно со стопроцентной вероятностью превратить в безошибочное решение;
- правильное решение будет наиболее вероятным, если лицу, принимающему решение, дополнительно предоставить информацию по стратегии, основанной на правиле (2).

Материал поступил в редакцию 20.06.2000.