ЧЕЛОВЕК И ТЕКСТ:

КАК И ЧТО МЫ ИЩЕМ, ПОЧЕМУ НАХОДИМ И НЕ НАХОДИМ

или о разнице между поиском в интернете стихотворения по стихотворной строчке и научной статьи по ее фрагменту, названию или по ключевым словам

М.Г. Крейнес

(ООО «БАЗИСНЫЕ ТЕХНОЛОГИИ)

Сложность поиска и способы ее уменьшения

Проблемы поиска

Модель поиска

Информационная технология КЛЮЧИ К ТЕКСТАМ®

Примеры

Сложность поиска

Определим сложность поиска как количество информации, соответствующей указанию источников, содержащих нужную информацию, среди множества всех найденных источников. Формулируя задачу таким образом, уйдем от попыток определить ценность и количество информации, содержащейся в конкретных источниках.

Предположим результативность поиска требуемых источников информации, выделим пять случаев – вся нужная информация содержится:

1) в одном единственном источнике из множества источников N_i, получаемого по единственному запросу i,

2) в любом источнике из некоторого подмножества M множества источников N_i , получаемого по единственному запросу i,

3) в некоторой совокупности L источников, когда каждый источник оказывается в составе множеств источников N_i, получаемых по отдельным запросам i, $i = 1,\ldots,L$,

4) в некоторой совокупности K источников из множества источников N_i , получаемого по единственному запросу i,

5) в некоторой совокупности наборов источников K_i, $i = 1,\ldots,G$, когда каждый из наборов источников оказывается в составе множеств источников N_i, получаемых по отдельным запросам i, $i = 1,\ldots,G$.

В первом случае необходимо выбрать единственный источник среди всего множества источников, которые потенциально могут ее содержать. Это соответствует количеству информации равному $\log_2 N_i$.

Во втором случае необходимо выбрать любой содержащий нужную информацию источник, что соответствует количеству информации, равному

$$\log_2 (N_i + 1 - M).$$

В третьем случае необходимо указать все источники из совокупности источников, содержащей требуемую информацию, что соответствует количеству информации, равному

$$\sum_{i=1}^{L} \log_2 N_i.$$

В четвертом случае необходимо указать все источники из совокупности источников, содержащей требуемую информацию. Предположим, что множество источников N_i рассматривается в последовательном ``многопроходном'' режиме с выявлением одного нужного источника за один проход и с исключением при очередном ``проходе'' ранее найденных источников, содержащих нужную информацию. Тогда, обозначив N_j – число источников, рассматриваемых на шаге j, получаем, что указание всех необходимых источников соответствует количеству информации

$$\sum_{j=1}^{K} \log_2 N_j = \sum_{j=1}^{K} \log_2 (N_i+1-j).$$

Пятый случай – комбинация третьего и четвертого случаев. Обозначив N_{ij} – число источников, рассматриваемых на шаге j шага i, получаем, что указание необходимого набора источников соответствует количеству информации

$$\sum_{j=1}^{G} \sum_{j=1}^{K_i} \log_2 N_{ij} = \sum_{j=1}^{G} \sum_{j=1}^{K_i} \log_2 (N_i +1-j).$$

Если в третьем и пятом случаях все N_i близки одной и той же величине N, в четвертом случае K много меньше N, и в пятом случае K_i, $i = 1,\ldots, G$ много меньше N, то можем объединить случаи 3 – 5. Указание F источников нужной информации во всех этих случаях будет примерно соответствовать количеству информации

$$F \log_2 N.$$

$$\log_2 (N_i + 1 - M)$$

$$F \log_2 N$$

Принципиальная разница сложности поиска стихотворения по стихотворной строчке и научной статьи по названию или по подготовленным автором ключевым словам:

- стихотворная строчка уникальна, ее словарный состав уникален, стихотворение приводится на большом числе сайтов,

- название научной статьи и ключевые слова в основном состоят из общеупотребительных научных терминов.

Как уменьшить сложность поиска - уменьшить число бессмысленных ответов поисковой системы?

- увеличить в результатах поиска долю текстов, содержащих нужную информацию

- научиться так формулировать запрос, чтобы необходимая информация содержалась в небольшом числе источников

- ранжировать результаты по соответствию запросу

Тривиальный ответ – искать в Википедии!

Корень проблем – несоответствие систем поиска информации, зафиксированной в форме текстов на естественных языках, требованиям пользователей

Надежен и прост поиск по прямым ссылкам на документ (библиографическим или в информационных сетях и базах данных), но где и как их найти по требованиям к содержанию документа?

Проблемы поиска

Основные причины неудовлетворительного качества систем поиска:

-1) ограниченность выразительных средств для формирования пользователем запросов, адекватно отражающих содержательные информационные потребности,

-2) неэффективность средств отбора информационных источников (содержательного сравнения запросов и текстов или описаний информационных источников) для поиска нужной пользователю информации,

-3) скудность аналитических средств, предназначенных для содержательного анализа и систематизации результатов поиска.

Принципиальные задачи создания систем поиска:

- разумная вербализация информационной потребности

- разумная организация поиска

- вычислительная объективная оценка результатов поиска

Модель поисковой системы

информационное хранилище - коллекция источников на естественных языках (например, традиционные ``бумажные'' библиотеки или источники информации в электронной форме)

зона поиска – описание информационного хранилища, технологически предназначенное для непосредственного выполнения процедур поиска

поисковая модель – модель зоны поиска, определяющая ее параметризацию и использование

зона поиска

- информационное хранилище, как оно есть,

- формализованные данные источников --- структурированная (в соответствии с традицией и правилами зоны поиска) информация, сопровождающая источники ``от рождения'' или с момента появления в зоне поиска (например, идентификатор источника, заглавие, год издания, название издания, автор, место работы авторов, местонахождение источника в хранилище --- библиотечный идентификатор источника или его адрес в информационной сети и т.д.),

- информация о содержании (семантике) источников информации и/или их коллекций,

- формальная несемантическая информация о содержании источников информации и/или их коллекций, характеризующая особенности текстов (их коллекций) как не интерпретируемых последовательностей символов (например, набор контрольных сумм для пересекающихся фрагментов текста одинакового размера, на которые разделен источник),

- информация об использовании отдельных источников и/или их коллекций (например, библиотечный или архивный формуляр источника, данные о ссылках на источник и в источнике),

- информация об использовании информационного хранилища в целом или его части (например, коллекция всех запросов, сформулированных при поиске информации в хранилище или в его части, соответствующей конкретным значениям метаинформации или определенным семантическим характеристикам),

- система поисковых индексов для организации поиска информации в зоне поиска.

поисковая модель

может включать в себя три функционально различных компоненты:

- модель представления информационного источника (статическую или динамическую),

- модель представления коллекции информационных источников (статическую или динамическую),

- модель отбора источников

полностью определяет информационное обеспечение и принципы функционирования систем сервисов

- по выполнению поиска в рамках зоны поиска и/или в части зоны поиска, соответствующей определенной коллекции источников информации,

- по формированию модели представления коллекции источников информации (например, для коллекции источников, полученной в результате поиска),

- по обработке результатов поиска, используемой для их анализа и предоставления пользователю.

поисковая модель

- формализованные модели источника,

- содержательные модели (модели семантики) источника информации и/или коллекций (наборов) источников информации,

- несодержательные модели (формальные модели) источника информации и/или коллекций (наборов) источников как не интерпретируемых содержательно последовательностей символов,

- модели прагматики для источника и/или коллекции источников, формализующие информацию о его использовании, например, ссылки на источник в других источниках информации, ссылки в источнике на другие источники информации (информационный контекст источника) или число запросов пользователями,

- модель запроса,

- критерии оценки приемлемости (адекватности) источника и/или коллекции источников для запроса на поиск информации.

Основные свойства (недостатки) поисковых моделей, реализованных в распространенных системах поиска

- компоненты формализованной модели источника должны быть известны пользователю априори,

- достоверные содержательные модели (модели семантики) источника информации и/или коллекций (наборов) источников информации отсутствуют (онтологии, модели, основанные на выявлении фактов, вероятностное моделирование тематической структуры коллекции текстов)

ПРИМЕР:

«У порога дома старика Батурина встретила супруга Петрова.»

- несодержательные модели (формальные модели) источника информации и/или коллекций (наборов) источников как не интерпретируемых содержательно последовательностей символов не позволяют сформировать содержательный запрос,

- модель запроса и формирование запроса не поддерживаются моделями семантики,

- критерии оценки адекватности источника и/или коллекции источников для запроса на поиск информации не основаны на достоверных моделях семантики.

Только модели прагматики нашли широкое применение в системах поиска:
- модели внутренней прагматики (ссылки в источнике),
- модели внешней прагматики (ссылки на источник и данные о его востребованности).

Поисковая модель в системах контекстного поиска использует тривиальные содержательные модели (модели семантики) источника информации в виде словарей слов источника или его определенных фрагментов.

Принципиальная разница сложности поиска стихотворения по стихотворной строчке и научной статьи по названию или по подготовленным автором ключевым словам:

- стихотворная строчка уникальна и ее словарный состав уникален,

- название научной статьи и ключевые слова в основном состоят из общеупотребительных научных терминов.

Поисковая модель, использующая нетривиальные модели семантики информационных источников и их коллекций, определяет выразительные средства для формулировки запроса и может обеспечивать помощь в формировании запроса и практически полезное ранжирование результатов

Информационная технология КЛЮЧИ К ТЕКСТАМ®
- технология использования контекстов слов информационных источников

Иерархия данных
- элементарный факт – наличие слова в тексте
- базовая переменная – число словоупотреблений слова в тексте
- исходные параметры моделей текстов и их коллекций:

** объем текста в числе словоупотреблений,

** объем коллекции текстов в числе словоупотреблений,

** частотный словарь текста (множество слов с указанием числа словоформ данного слова, встретившихся в тексте),

** частотный словарь коллекции текстов (множество слов с указанием числа словоформ данного слова, встретившихся во всех текстах коллекции),

** число текстов коллекции, в которых встречается определенный набор слов.

Базовая гипотеза формирования моделей семантики: основным носителем семантики произвольного текста на естественном языке является множество слов, отличающее текст от множества текстов, определяющих естественный язык, на котором написан текст. Их определение основано на выявлении пар слов, устойчиво связанных в конкретном тексте и не демонстрирующих такой связи в множестве текстов, определяющем естественный язык, на котором написан текст.

Критерий выбора слов, представляющих модель семантики: комбинаторный индекс, который рассчитывается для каждой пары слов в тексте по числу словоупотреблений каждого из слов пары в тексте и множестве текстов, определяющем естественный язык, на котором написан текст, с учетом объемов текста и множества текстов. Такой индекс устойчивости связи слов не зависит от грамматической сочетаемости слов, от взаимного расположения слов в тексте и от любой содержательной информации о конкретных словах и их сочетаниях.

Модели содержания текстов (статические)
- **верно отражают тематику и содержание текста**
- **воспроизводимы**
- **лаконичны**
- **интерпретируемы**
- **технологичны**

Модели содержания текстовых коллекций (динамические), формируемые в информационной технологии КЛЮЧИ К ТЕКСТАМ®:

А) групповая модель, формируемая в результате вычислительного поиска близких по содержанию документов (кластеризации);

Б) словарные модели, формируемые в результате анализа моделей текстов коллекции.

Словарные модели коллекции:

- неструктурированная модель, характеризующая суммарную значимость слов для представления содержания коллекции;

- оптимизированная модель (система терминов), описывающая коллекцию с требуемой пользователем полнотой двумя группами слов (одна характеризует коллекцию в целом – стилистические детерминанты, вторая позволяет выделить в коллекции отдельные тематические группы – семантические детерминанты);

- структурная модель, описывающая тематические группы документов коллекции (используется как критерий остановки уточнения оптимизированной модели).

Поисковая модель информационной технологии КЛЮЧИ К ТЕКСТАМ®:

- статические модели семантики текстов

- динамические модели семантики текстовых коллекций

- отбор и анализ результатов поиска на основе анализа моделей семантики

- помощь в формировании запроса в форме списка слов

- технологическое обеспечение «понимания» системой запроса в форме текста на естественном языке

Пример неструктурированной модели

Всего общих слов: 655481

УНИАН 146813 руб 384655 проц 94705 Украины 207156 Госдумы 91063 совещании 56145 долл 185745 ПРАЙМ 167827 ветеранов 52601 Известия 73938 водитель 49067 выразил 64364 банка 218033 млн 279643 Ирака 37637 налога 166294 акций 188497 сельского 66739 утвердить 46019 Нижегородской 43072 обсудить 41845 пенсии 52109 ЛУКОЙЛ 29625 департамента 85962 млрд 150393 ИНТЕРФАКС 96080 коллектив 75667 примет 69809 марки 67248 поставки 71028 компании 334190 выиграл 47510 индекс 27527 бирже 48427 активов 47811 пострадавших 41959 спорта 60566 чемпионата 63629 кодекса 60780 автомобилей 103163 соревнований 43067 управляющего 38091 фестиваля 42660 РАО 41885 законопроект 49085 торговли 215038 сокращение 62857 ремонт 63278 топлива 54691 пенсионеров 43290 займа 52814 Газпром 43865 сезона 69239 валюты 63062 Максим 42650 транспорта 72979 энергетической 47122 кредитных 56068 нефти 77824 завершить 73711 турнира 38557 медали 33606 губернатора 116456 погашения 61724 муниципальных 61570 предстоящих 47490 евро 44564 инвалидов 38004 иск 43010 Источник 280524 рассмотреть 37339 продолжить 51850 визита 69911 сырья 64937 нефтяной 58327 экспорта 56348 пожара 29531 студентов 52690 приз 45425 лицензии 52333 тепло 75353 депутатов 165018 Путин 74906 тыс 179417 возросла 41209 предусматривает 69584 облигаций 30489 производителей 79922 Минфина 43597 транспортных 64364 украинских 90319 сообщил 453468 направить 70510 двигателя 33770 задолженности 80465 пострадали 37084 задержан 44189 москвичей 52767 Белоруссии 37015 спортивных 61608 потребителей 63429 полномочий 53192 добычи 42007 белорусских 24786 погибли 56201 РФ 308856 исполнения 79620 конкурса 89543 областной 126613 сделки 49687 фракции 37151 тонн 89682 предприятий 335909 жилищно 46987 приступить 49386 информационных 84192 края 125504 выполнить 49910 высказал 53276 области 376530 наркотиков 29418 превышает 62644 обслуживания 61117 заседании 165002 грн 43279 пассажиров 40002 поселке 49494 бюджета 202833 выпуска 191485 акционеров 65403 размещения 46543 заболевания 35166 школы 135681 подчеркнул 108702 Людмила 33816 избирателей 38819 превысил 36615 экологической 37607 лечения 35113 спектакль 35198 металлов 45656 исполнительной 69076 учебных 52369 учителя 43779 постановление 105374 творческих 51516 Законодательного 79277 девять 51145 районного 56426 Красноярского 33201 стандартам 46271 филиала 44208 номинала 21632 чемпион 41766 договоренности 39269 предоставления 79007 тур 36751 подать 58979 Сибирский 40284 приобретение 60239 утверждении 84918 инвестиционных 73918 реконструкции 51813 внести 62232 старт 42999 выпустить 35118 военнослужащих 32247 Олимпийских 28558 морской 48157 задержали 34019 чтении 40478 округа 107529 распоряжение 64296 выборов 172859 сборной 43758 таможенных 41588 кино 51109 изъято 35623 стратегических 52608 завод 158173 организаторы 59444 пятницу 72969 комбинат 40500 ограничения 60006 АО 67100 палаты 62897 сократить 68878 директор 263183 повысить 45891 осуществления 60849 поможет 32271 женской 41561 посмотреть 40261 предпринимателей 72818 составил 359560 завершения 60553 очков 31727 больницы 62654 ребенка 45673 резервов 37924

Пример оптимизированной модели (системы терминов)

Семантические детерминанты коллекции

франк; вручил; номинала; МВФ; Швейцарский; терактов; проголосовали; космической; белорусских; лауреат; посетителей; Уральского; произошел; пособия; призвал; потребительских; назначил; приговор; учеников; индекс; строителей; аукционе; поезда; террористов; Олимпийских; месторождения; ЦИК; наркотиков; пожара; ЛУКОЙЛ; инспектор; облигаций; Сбербанк; претендентов; арбитражный; кредиторов; аэропорта; очков; маршрут; проведет; принадлежащих; поможет; операторов; КУЧМА; сооружений; Урал; способствовать; ЦБ; зимний; Самарской; футбол; Красноярского; медали; юбилей; Новосибирск; двигателя; Людмила; Лужков; задержали; памятник; полугодии; автобус; МИД; ЕЭС; финале; наказания; билеты; лечения; разделе; заболевания; спектакль; голосования; изъято; обучения; итальянской; дополнений; соперников; заключить; взрыва; заработать; превысил; тур; скорость; милиционеры; масла; окружающей; пострадали; совершил; Немцов; фракции; ООН; избран; рассмотреть; аналитики; погибших; Всероссийского; экологической; Ирака; инвалидов; резервов; управляющего; менеджеров; британской; господин; домашних; турнира; избирателей; Вашингтон; заведения; Краснодарского; рейтинг; тренер; актер; договоренности; японских; артист; корпорации; пассажиров; Сибирский; посмотреть; железнодорожного; чтении; комбинат; Новгород; электроэнергии; женской; км; обсудить; чемпион; пострадавших; РАО; упал; добычи; конкурентов; награды; фестиваля; Максим; иск; старт; Нижегородской; соревнований; пенсионеров; грн; премии; Минфина; учителя; сборной; Газпром; принадлежит; задержан; филиала; падение; изделий; организовать; евро; художника; приз; девушки; металлов; ребенка; повысить; утвердить; избирательной; болезни; продавать; размещения; жилищно; энергетической; концерт; выиграл; строительных; природного; активов; энергии; вклад; морской; бирже; тарифов; достиг; миллиардов; льготы; водитель; законопроект; берегу; приступить; сделки; парламентских; сыграть; расследование; кино; выделено; творческих; продолжить; реконструкции; пенсии; правоохранительных; учебных; ветеранов; матч; существующих; студентов; стратегических; займа; молодежи; музея; устанавливается; полномочий; энергетики; высказал; достигнута; предоставить; топлива; заказ; кредитных; сад; погибли; экспорта; районного; предстоит; французской; предоставляет; Челябинской; нефтяной; мощности; категории; подать; победителей; прогноз; организаторы; поддержать; ограничения; Наталья; приобретение; спорта; обнаружили; привлечения; кодекса; осуществления; памяти; обслуживания; оперативно; погашения; муниципальных; спортивных; внести; обсуждения; превышает; больницы; палаты; покупки; валюты; ремонт; потребителей; обязанности; чемпионата; внесении; распоряжение; выразил; транспортных; СМИ; Верховной; установленном; сырья; акционеров; Нижнем; субъектов; родителей; марки; театра; исполнительной; сократить; сезона; примет; предусматривает; ценных; визита; подписания; старший; поставки; понедельник; пятницу; транспорта; завершить; редакции; инвестиционных; тепло; коллектив; подписал; нефти; Законодательного; предоставления; серии; сбора; исполнения; обсуждать; производителей; установить; задолженности; кандидатов; праздник; утверждении; отрасли; конкурса; тонн; украинских; возрасте; Госдумы; определения; Санкт; проц; мэра; ИНТЕРФАКС; установлены; кредитов; автомобилей; текущего; постановление; (всего слов: 317)

Стилистические детерминанты коллекции

протокол; верхней; военнослужащих; штраф; выпустить; республиканского; Наименование; Белоруссии; разработать; компьютерной; телефонной; Кореи; возросла; таможенных; поручено; уставного; стандартам; Всемирного; предстоящих; поселке; выполнить; баланса; девять; лицензии; москвичей; отменить; указ; совещании; оборот; подразделения; осенью; завершения; работающих; сокращение; сельского; АО; подтвердил; формирования; проверки; направить; предпринимателей; Южной; Известия; Путин; информационных; департамента; округа; подчеркнул; хозяйства; курс; губернатора; товаров; партии; края; агентства; областной; долга; собрания; увеличить; работников; школы; доходов; прибыль; газа; планируется; сообщает; УНИАН; млрд; поступления; летний; завод; американских; Федерации; Киев; депутатов; заседании; налога; ПРАЙМ; детей; процентов; выборов; тыс; городской; долл; акций; долларов; выпуска; бюджета; Украины; торговли; фонда; банка; суд; объем; администрации; района; закона; Цена; министр; директор; млн; Источник; РФ; компании; предприятий; составил; области; руб; сообщил; (всего слов: 109)

Пример структурной модели

все слова:

426 покрыли 97% док.

гр.1 (слов - 55, док - 295281) проголосовали, депутатов, Госдумы, парламентских, фракции

гр.2 (слов - 52, док - 304710) банка, займа, кредитных, Минфина, облигаций

гр.3 (слов - 45, док - 332695) театра, творческих, художника, актер, спектакль

гр.4 (слов - 43, док - 306275) турнира, сборной, чемпионата, чемпион, тренер

гр.5 (слов - 41, док - 176098) возросла, увеличить, сократить, составил, проц

гр.6 (слов - 36, док - 264887) газа, энергетической, добычи, энергетики, поставки

гр.7 (слов - 33, док - 227112) детей, родителей, ребенка, школы, учителя

гр.8 (слов - 30, док - 87346) проц, возросла, увеличить, составил, полугодии

гр.9 (слов - 28, док - 176655) закона, законопроект, чтении, внести, Законодательного

гр.10 (слов - 27, док - 157915) МИД, министр, ООН, визита, подчеркнул

гр.11 (слов - 25, док - 193085) пассажиров, маршрут, транспортных, автобус, км

гр.12 (слов - 23, док - 184114) задержан, правоохранительных, задержали, изъято, милиционеры

гр.13 (слов - 16, док - 74441) иск, суд, арбитражный, подать, управляющего

гр.14 (слов - 15, док - 86768) ЕЭС, РАО, ЛУКОЙЛ, Газпром, акций

гр.15 (слов - 14, док - 103267) установленном, определения, устанавливается, установлены, осуществления

гр.16 (слов - 13, док - 109831) произошел, пострадавших, погибли, пострадали, погибших

гр.17 (слов - 12, док - 79975) франк, Швейцарский, евро, японских, курс

гр.18 (слов - 12, док - 59113) строительных, реконструкции, сооружений, строителей, городской

гр.19 (слов - 12, док - 66548) пособия, инвалидов, пенсионеров, пенсии, работников

гр.20 (слов - 11, док - 66295) сырья, тонн, изделий, экспорта, металлов

гр.21 (слов - 11, док - 49747) корпорации, менеджеров, компании, стратегических, стандартам

гр.22 (слов - 11, док - 69261) Уральского, Новосибирск, области, Красноярского, Челябинской

гр.23 (слов - 8, док - 46414) упал, падение, индекс, аналитики, продолжить

гр.24 (слов - 8, док - 35379) Ирака, Вашингтон, ООН, британской, американских

гр.25 (слов - 7, док - 47903) УНИАН, Украины, Киев, КУЧМА, Верховной

гр.26 (слов - 6, док - 39619) исполнительной, обязанности, исполнения, полномочий, поручено

гр.27 (слов - 6, док - 21540) края, Краснодарского, администрации, сельского, районного

гр.28 (слов - 5, док - 21480) бирже, торговли, индекс, покупки, оборот

гр.29 (слов - 5, док - 10645) поселке, района, администрации, районного, сельского

гр.30 (слов - 4, док - 18841) акционеров, принадлежит, АО, директор

гр.31 (слов - 4, док - 4781) проверки, штраф, протокол, инспектор

гр.32 (слов - 4, док - 15519) завод, комбинат, предприятий, металлов

гр.33 (слов - 4, док - 26192) Нижегородской, Новгород, Самарской, губернатора

гр.34 (слов - 3, док - 23997) постановление, распоряжение, поручено

гр.35 (слов - 3, док - 28324) понедельник, пятницу, поступления

гр.36 (слов - 3, док - 22261) ремонт, жилищно, хозяйства

гр.37 (слов - 3, док - 19956) мэра, Лужков, москвичей

гр.38 (слов - 2, док - 24644) подписания, подписал

гр.39 (слов - 2, док - 19602) предоставляет, предоставления

гр.40 (слов - 2, док - 10473) налога, сбора

гр.41 (слов - 2, док - 14626) категории, льготы

Пример моделей семантики коллекции

для журнала Science

«ключевые слова»

Годы публикации	Слова модели, упорядоченные по неубыванию суммарного веса в коллекции, число документов, в которых слово встречается
1880-1909	Professor 12065 scientific 9488 species 5894 experiments 6499 observations 6921 paper 10432 author 7456 laboratory 4974 College 6536 Society 8841 volume 5623 chemical 4652 birds 2865 Academy 4386 subject 10985 inches 3497
1910-1939	Professor 13634 scientific 13904 laboratory 11984 experiments 9447 College 11045 species 5757 chemical 7953 medical 8057 Society 10883 Institute 8318 chemistry 7636 Medicine 5825 lectures 4710 Association 9026 Academy 5270
1940-1959	scientific 7986 laboratory 10027 experiments 7576 Professor 4371 Institute 7087 chemical 6705 Medicine 5314 York 8562 medical 6182 College 6831 chemistry 5546 Chicago 4099 scientists 4111 physics 4051 Society 5781 Department 9949
1960-1979	scientific 12204 experiments 15319 laboratory 17845 scientists 10482 percent 19027 York 20441 concentration 10489 species 8683 volume 12104 chapter 4059 chemical 11090 paper 11548 rats 5570 book 9901 illus 5659 observed 14845
1980-1996	scientific 11958 scientists 11753 DNA 8324 researchers 9470 experiments 14540 laboratory 15915 species 8732 genetic 7942 Institute 17071 mice 5180 binding 7197 cells 15518 human 15090 paper 8821 chemical 9215 concentration 9401

система терминов для коллекций материалов журнала Science, сгруппированных по времени публикации

Годы публикации	Слова, представляющие семантические детерминанты коллекции и упорядоченные по неубыванию веса
1880-1909	nerve; storm; cloud; seed; wire; instructor; railway; shell; tribes; bone; steam; psychology; stars; eggs; meetings; catalogue; coal; ray; wave; angle; vessel; tube; metal; marine; philosophy; Louis; ice; sand; forest; mathematics; depth; gas; mental; organic;
1910-1939	virus; mice; bone; crystal; coal; strain; alcohol; male; female; slide; sugar; map; rats; medium; genetic; fruit; root; edition; seed; marine; eggs; iron; instructor; metal; concentration; honorary; wave; Pacific; deposited; medal; Canada; motion; tube;
1940-1959	nerve; slide; illus; muscle; seed; eggs; root; female; diet; birds; rubber; chamber; crystal; medal; virus; inches; serum; liver; alcohol; edition; fellowship; mice; exposure; AAAS; negative; patient; dose; particles; vitamin; elected; dean; count; reader;
1960-1979	antibody; Senate; virus; edition; birds; RNA; nerve; sediments; Soviet; hormone; pollution; editorial; muscle; panel; DNA; solar; soil; ocean; visual; mathematics; crystal; educational; rocks; serum; assistant; session; oxygen; marine; Canada; female; patient;
1980-1996	advertising; Ca2; HIV; Senate; prize; weapons; FAX; cloud; bone; software; Mar; therapy; NASA; Soviet; editorial; muscle; satellite; laser; emission; chapter; chromosome; climate; marine; infected; ice; neurons; solar; ocean;

структурированная модель (публикации 1980- 1996 годы)

№ группы (Число текстов в группе)	Словарный состав элемента структурированной уточненной модели (вес – значимость слова в группе)
1 (15190)	Семантические детерминанты fragment (99%), mutant (99%), chain (99%), amino (99%), probe (99%), residues (99%), binding (99%), domain (99%), peptide (99%), RNA (99%), enzyme (99%), DNA (99%), mm (98%), gene (98%), expression (98%), mutations (98%), membrane (97%), lane (94%), receptor (94%), regulation (92%), Ca2 (87%), induced (87%), mice (85%), muscle (84%), neurons (84%), strain (81%), culture (79%), antigen (78%), tissue (77%), chromosome (77%), rats (77%), antibody (77%), molecules (76%), active (75%), procedure (68%), brain (68%), concentration (67%), genetic (66%), column (66%), normal (64%), Biology (64%), map (61%), growth (55%), interaction (50%), cycle (49%), formation (40%), production (34%), (всего слов: 47) Стилистические детерминанты protein (99%), molecular (99%), indicated (99%), site (99%), mechanism (87%), analysis (84%), experiments (79%), complex (76%), factor (63%), indicate (61%), cells (51%), study (40%), occur (40%), circle (37%), observed (35%), percent (35%), Department (33%), proposed (31%), observations (25%) (всего слов: 19)
2 (10295)	Семантические детерминанты federal (100%), committee (99%), budget (99%), funds (99%), Congress (99%), Senate (99%), agencies (99%), project (99%), administration (99%), facility (99%), weapons (99%), panel (96%), resources (93%), comments (92%), News (90%), Foundation (85%), editorial (82%), advertising (82%), Academy (81%), assistant (73%), FAX (69%), Japan (69%), associate (65%), institutions (60%), Soviet (58%), NASA (50%), technical (50%), June (45%), environmental (45%), basic (45%), production (41%), instrument (34%), (всего слов: 32) Стилистические детерминанты program (99%), National (99%), scientists (99%), scientific (77%), Washington (64%), grant (63%), York (46%), volume (34%), circle (28%) (всего слов: 9)
3 (8013)	Семантические детерминанты ocean (97%), global (91%), cloud (81%), earth (78%), Mar (76%), zone (74%), estimate (73%), satellite (72%), solar (72%), climate (67%), ice

(63%), marine (56%), density (54%), carbon (53%), NASA (53%), mass (50%), impact (50%), California (50%), emission (50%), natural (49%), distribution (49%), constant (48%), resolution (46%), origin (46%), environmental (44%), core (44%), wave (42%), environment (37%), error (37%), evolution (37%), formation (33%), conclusion (26%) (всего слов: 32)

Стилистические детерминанты

techniques (55%), model (52%), laboratory (43%), center (40%), measured (35%), observations (33%), percent (30%), behavior (30%), proposed (28%), (всего слов: 9)

4 (7278)

Семантические детерминанты

metal (94%), atoms (94%), solid (94%), laser (87%), electron (86%), magnetic (83%), chemical (79%), chemistry (72%), ray (67%), crystal (65%), liquid (63%), ion (60%), transition (60%), density (59%), constant (59%), emission (58%), peak (57%), resolution (56%), wave (56%), organic (55%), carbon (53%), natural (49%), distribution (49%), mass (45%), core (44%), environment (37%), error (37%), instrument (34%), formation (33%), conclusion (26%) (всего слов: 30)

Стилистические детерминанты

techniques (55%), model (52%), York (46%), laboratory (45%), measured (43%), nuclear (42%), center (40%), circle (35%), behavior (35%), observations (33%), percent (30%), volume (30%), proposed (28%), (всего слов: 13)

5 (6119)

Семантические детерминанты

patient (100%), disease (100%), cancer (100%), medical (99%), therapy (99%), HIV (99%), aid (99%), clinical (99%), infection (99%), blood (99%), drug (99%), infected (99%), bone (98%), virus (97%), tumor (96%), Medicine (94%), animals (68%), health (66%), NIH (61%), population (42%), production (33%), (всего слов: 21)

Стилистические детерминанты

human (99%), Institute (96%), developed (47%), study (39%), report (33%), development (33%), Department (26%) (всего слов: 7)

6 (5860)

Семантические детерминанты

book (98%), reader (97%), discussion (95%), author (92%), Chicago (86%), advertising (85%), editorial (84%), chapter (82%), paper (78%), FAX (76%), illus (75%), associate (68%), AAAS (68%), publication (63%), prize (61%), basic (45%), Academic (40%), instrument (34%) (всего слов: 18)

Стилистические детерминанты

York (55%), volume (40%), (всего слов: 2)

7 (2174)

Семантические детерминанты

	physics (75%), particles (75%), density (54%), carbon (53%), emission (49%), natural (49%), distribution (49%), constant (48%), wave (47%), resolution (46%), mass (46%), core (42%), environment (37%), error (37%), conclusion (26%) (всего слов: 15)
	Стилистические детерминанты
	energy (79%), techniques (55%), model (52%), laboratory (49%), center (40%), measured (35%), (всего слов: 6)
8 (2133)	Семантические детерминанты
	animals (90%), male (84%), brain (83%), female (81%), neurons (69%), muscle (63%), population (52%), species (47%), (всего слов: 8)
	Стилистические детерминанты
	report (36%), development (30%), behavior (29%) (всего слов: 3)
9 (1021)	Семантические детерминанты
	students (86%), graduate (77%), social (62%), Academic (50%), basic (49%) (всего слов: 5)
10 (1001)	Семантические детерминанты
	lab (71%), researchers (71%), technical (42%)
11 (845)	Семантические детерминанты
	dose (86%), animals (81%), brain (76%), neurons (69%), muscle (63%), exposure (62%), (всего слов: 6)
	Стилистические детерминанты
	study (44%), report (44%), Department (33%)

Выбор критерия остановки формирования системы терминов (оптимизированной модели семантики коллекции)

Модель коллекции текстов (всего – 472390 текстов) построена по		1000 слов	800 словам	600 словам	400 словам
1	Число текстов, категоризованных по семантическим детерминантам	470571	470310	469528	467032
2	Число тематических категорий в модели	62	33	20	17
3	Среднее притяжение	0.532	0. 520	0. 443	0.376
4	Среднее максимальное притяжение	0.63	0.609	0. 535	0.458
5	Число категоризованных текстов	470753	470599	470135	468881
6	Число текстов, категоризованных только по стилистическим детерминантам	2	289	607	1849
7	В том числе с притяжением более 10%	461790	457831	438656	408383
8	В том числе с притяжением более 20%	448349	444617	425445	391614
	В том числе с притяжением менее 20 %	22404	25982	44690	77267

ТЕКСТЫ И АННОТАЦИИ

Модель построена по коллекции		Текстов (всего – 472390)	Аннотаций (всего – 473899)	Аннотаций (всего – 473899)	Текстов (всего - 472390)
№	Характеристика категоризации	Коллекция текстов	Коллекция аннотаций	Коллекция текстов	Коллекция аннотаций
1	Число тематических категорий в модели	62	58	58	62
2	Среднее притяжение	0.532	0.304	0.392	0.223
3	Среднее максимальное притяжение	0.63	0.42	0.468	0.358
4	Число категоризованных документов	470753	473370	468580	473296
5	В том числе с притяжением более 10%	461790	434893	424803	400492
6	В том числе с притяжением более 20%	448349	392631	396809	334987
7	Число пар текст/аннотация, попавших в сопоставимые категории	373266		380403	376234
8	Число пар текст/аннотация, попавших в сопоставимые категории по максимальному притяжению	258568		248994	274252
9	Число пар текст/аннотация, попавших в сопоставимые категории для первых 9 категорий	361485		357362	348225
10	Число пар текст/аннотация, попавших в сопоставимые категории для первых 20 категорий	368121		375816	365288

Реализованная в информационной технологии

КЛЮЧИ К ТЕКСТАМ® поисковая модель основана на достоверных моделях семантики текстов и их коллекций и обеспечивает

-1) разнообразие естественных для человека выразительных средств для формирования пользователем запросов, адекватно отражающих содержательные информационные потребности (слова, слова – основные носители содержания текста, тексты произвольного размера),

-2) эффективность средств отбора информационных источников (содержательного сравнения запросов и текстов или описаний информационных источников) для поиска нужной пользователю информации,

-3) разнообразие аналитических средств, предназначенных для содержательного анализа и систематизации результатов поиска.