УДК 615.035.4

Ауезова А.С. [1], Муратова К.Н. [1], Синчев Б. [1]

[1] Международный университет информационных технологий, Алматы, Казахстан

МЕТОДЫ ИНФОРМАЦИОННОГО ПОИСКА НЕСТРУКТУРИРОВАННЫХ ДАННЫХ

Аннотация. *В статье предложен новый метод, применяемый для решения задачи информационного поиска неструктурированных (текстовых) данных. Поиск документов осуществляется по ключевым словам, на естественном языке, применяемым в поисковых машинах. Таким образом, на основе полиномиальных алгоритмов создается универсальная машина выборки по нескольким ключам с улучшенными характеристиками по времени и пространству. Данная предлагаемая машина может быть применена для обработки больших данных в различных областях экономики. Для достижения цели в основе новых полиномиальных алгоритмов использована задача о сумме подмножеств, которая относится к классу NP-complete. Эти алгоритмы значительно эффективнее по времени и пространству существующих лучших полиномиальных и экспоненциальных алгоритмов.*

Ключевые слова*: поиск, метод, алгоритм, неструктурированная информация, поисковая машина, полиномиальные алгоритмы, NP-complete, большие данные.*

Введение

В настоящее время существуют важные практические и теоретические задачи: теория алгоритмов (theory of algorithms), задача поиска (search problem), задача выполнимости (satisfiability problem), задача принятия решений (decision problem), задача кодирования (encryption problem), задача шифрования (encipherment problem) и другие. Одним из ключевых мест является анализ времени работы алгоритмов и используемая ими память. Задача нахождения k-мерного подмножества из n-мерного множества, сумма элементов которого равна некоторому заданному числу S, является NP-полной (NP-complete). Вычислительная сложность этой задачи зависит от двух параметров – количества n элементов исходного множества и точности p (определяется как количество двоичных разрядов в числах, составляющих множество). Задача становится лёгкой только при очень малых значениях параметров n и p. Если n (количество входных данных) мало, то полный перебор вполне приемлем. Если параметр p (количество разрядов в числах множества) мал, можно использовать динамическое программирование при решении задачи о сумме подмножеств. В классических работах определены время работы алгоритма $T=O(2^{n/2})$ и потребляемая память $M=O(2^{n/4})$, которые не позволяют применять полученные результаты на практике при больших значениях n и на поиск решения необходимо затратить экспоненциальное время. Основным недостатком известных табличных методов (сумм) является построение каждой строки таблицы по свойству определяемым каждым ключевым словом.

К примеру, одна из самых крупных поисковых систем Google рассматривает более 200 различных показателей при оценке веб-сайтов, включая текст, внутренние ссылки, удобство использования веб-сайта и информационную архитектуру. Чтобы получить и сохранить долю рынка онлайн-поиска, поисковым системам необходимо убедиться, что они предоставляют результаты, соответствующие тому, что ищут их пользователи. Они делают это, поддерживая базы данных веб-страниц, которые они разрабатывают с помощью автоматизированных программ, известных как «пауки» или «роботы» для сбора информации. Поисковые системы используют сложные алгоритмы для оценки веб-сайтов и веб-страниц и присваивают им рейтинг по релевантным поисковым фразам. Эти алгоритмы тщательно охраняются и часто обновляются. Поскольку потребители и организации все больше полагаются на поисковые системы при выборке товаров, услуг и поставщиков, в которых они нуждаются, важность поисковых машин с меньшими затратами для современного бизнеса только возрастает.

Методы сбора первичной (исходной) информации

Среди существующих поисковых систем самыми популярными являются два принципа информационного поиска: поиск по ключевым терминам и поиск на основе кластерных методов и векторных моделей [1],[9]. В большинстве случаев информационно-поисковые системы используют метод поиска по ключевым словам, как основной, а кластерные методы в качестве дополнения [10], что существенно повышает эффективность и точность поиска информации. В системах, использующих поиск по ключевым словам, как Google, Яндекс, AstaVista, Yahoo, процесс поиска сводится к поиску во множестве документов вхождений каждого из заданных ключевых слов, с последующей сортировкой этих документов по степени релевантности. В подобных системах процесс поиска сводится к поиску во множестве документов вхождений каждого из заданных ключевых слов, с последующей сортировкой этих документов по степени релевантности.

Приведем методы информационного поиска: последовательный поиск, точный поиск по алгоритму Бойера-Мура, информационно-поисковые системы, строящие поисковый индекс, хеширующие методы, инвертированные файлы, сигнатурные файлы, суффиксные массивы, деревья, суффиксные деревья, кластерные методы, тернарные деревья поиска и векторные модели, триангуляционные деревья и др.

Эти подходы позволяют перейти к самым современным методам информационного поиска.

Совершенствование работы поисковых систем невозможно без сбора достоверной информации и последующего ее анализа. Поэтому в общей теории и практике информационного поиска повышенное внимание уделяется теоретическим исследованиям, включающим в себя методы сбора информации и ее анализа.

Теоретические исследования используют научную базу таких дисциплин как теория алгоритмов, теория поиска, теория шифрования, теория кодирования, теория принятия решений и др. К основным методам, применяемым для решения научных и практических задач следует отнести:

– технология обработки и анализа больших данных;

– методы исследования вычислительных операций;

– методы алгоритмической разрешимости.

В рамках совершенствования поисковых систем пристальное внимание уделяется дискретным оптимизационным задачам в комбинаторике, такие как задача о ранце.

Описание основных научных вопросов и гипотез

Вопрос о равноправии классов задач сложности *P* и *NP* (также общеизвестный как проблема перебора) — это одна из важных открытых проблем теории алгоритмов уже больше чем пяти десятилетий. Если задачи сложности *P* и *NP* получить утвердительный ответ, это будет означать, что теоретически возможно решать многие сложные задачи существенно быстрее, чем сейчас.

В разделе теории о вычислительной сложности алгоритмов рассматриваются отношения между классами P и NP. Она изучает наиболее общие ресурсы для решения некоторой задачи. Ресурсы — это время (количество шагов) и память (необходимый объем памяти для решения задачи). Проблема равенства классов P и NP входит в семь задач тысячелетия, за решение которой Математический институт Клэя назначил премию в миллион долларов США.

Теоретические и прикладные исследования проблемы базируются на методах системного анализа, Web Service, Big Data, Hadoop/Map Reduce.

Машина, созданная по данной методике, базируется на основные положения по разработке поисковых систем, а доступ к ним осуществляется на основе поискового запроса на естественном языке.

Реализация машины осуществляется на основе современной теории класса NP-complete.

Исследование функциональных качеств машины проводится на основе контрольных баз данных, а корректность работы проверяется и доказывается методами и полиномиальными алгоритмами. Результаты научного исследования по данному проекту позволят:

- развивать теорию алгоритмов и практику поиска произвольной информации;
- решить сложные задачи, связанные с равенством классов P и NP;
- подготовить специалистов и молодых ученых в области информационных технологий;
- создать технологии и инструментарии для разработки поисковых машин.

На сегодняшний день имеются компании, которые заинтересованы в результатах данного проекта. Это компании, которые занимаются большими данными.

Обоснование научной новизны

В n-элементном множестве целых чисел найти подмножество, сумма элементов которого близка сертификату S. Эта задача о сумме подмножеств связана с проблемой ранца, которая решена Р. Беллманом [1] в 1956 году. Им предложен псевдополиномиальный алгоритм с временем выполнения $T=O(nS)$ на основе метода динамического программирования. Поскольку этот алгоритм псевдополиномиального времени является фундаментальной частью информационных технологий, а Subset Sum является одной из основных NP-трудных проблем, важно решить, можно ли улучшить время выполнения $O(nS)$.

Спустя 43 года Д. Писинжер [2] улучшил решение этой задачи за время $T=O(nS/logS)$, применяя битовое представление сертификата S.

Недавно Килильяс и Сюй представили в [3] псевдополиномиальный алгоритм «разделяй и властвуй», который вычисляет все реализуемые суммы до целого числа в оценке времени O (min(, , σ)) выполнения алгоритма, где σ - сумма всех элементов исходного множества. Они используют хеширование для быстрого решения задачи о сумме подмножеств, если исходное множество содержится в небольшом интервале. Затем они сводят общий случай к случаю малого интервала путем расщепления исходного множества на меньшие подмножества. Они считают, что новый алгоритм является самым быстрым общим детерминистическим алгоритмом для этой задачи. Однако решение исходной задачи находится с некоторой точностью.

В 2017 году К. Биргман предложил псевдополиномиальный рандомизированный алгоритм [4], работающий за время $O(n+S)$. Точнее, учитывая экземпляр задачи о сумме подмножеств (, S), вычисляется множество всех сумм $s(; S)$, генерируемых небольшими подмножествами $Y \subseteq$ размерности $|Y| \leq k$ с постоянной вероятностью принадлежности сертификата S множеству сумм $s(; t)$. В работе не указана величина вероятности.

Важно отметить, что детальный обзор современных результатов, содержащихся в более 60 научных трудах, по известным лучшим алгоритмам решения задачи о сумме подмножеств приведен в работах [3, 4].

Наряду с псевдополиномиальными алгоритмами имеются точные алгоритмы с экспоненциальным временем работы, когда в n-элементном множестве существует хотя одно подмножество, сумма элементов которого равна S. В первую очередь, к ним относятся работы Горовиц, Санни [5] с временем выполнения алгоритма и требуемым пространством и Шреппеля, Шамира [6] с временем выполнения алгоритма и требуемым пространством При этом они использовали инвариантные соотношения.

В работах [7,8,9] введены понятия полиномиального (практического) алгоритма и его сложности. В [10] показано, что сложность работы алгоритма для задачи о сумме подмножеств составляет не менее $2^{n/2}$, где n — число переменных. Более детальное исследование сложности решения задачи о сумме подмножеств проведено в работах [11,12]. В них полу-

чены верхние и нижние оценки максимальной экспоненциальной сложности решения этой задачи методом ветвей и границ.

Сложность алгоритма напрямую используется при определении времени работы известных алгоритмов поискового запроса неструктурированной информации по многим ключевым словам [13,14] и методов обработки и анализа больших данных [15].

Важно подчеркнуть, что определение сложности решения задачи о сумме подмножеств является фундаментальной проблемой, используемой в качестве стандартной задачи, которая может быть решена за слабо полиномиальное время во многих дисциплинах по образовательным программам. В качестве основной задачи в теоретической информатике и учебных материалах [16,17] рассматривается задача *k*-SUM, которая является подзадачей задачи о сумме подмножеств и связана с методом перебора.

В работе [18] для параметра мощности k=2 разработаны уникальный метод и универсальные алгоритмы с временем и требуемым пространством для выборки искомого подмножества и для параметра мощности k=3. *T*. Для параметра сложности k=2 показана разрешимость равенства классов P и NP. Другими словами, найдено решение этой задачи, в которой проверка правильности решения задачи будет равна получению решения.

Описание методов исследования

Основные методы исследования базируются на современных методах и алгоритмах решения базовой задачи-задача о сумме подмножеств, которая относится к классу NP-complete. Решение хотя бы одной задачи из этого класса позволяет решить все оставшиеся задачи этого класса.

Постановка задачи. Задача о сумме подмножеств формулируется в виде:

$$\sum_{i=1}^{n} \quad a_i x_i = S, a_i \in \{0,1\}, x_i \in X^n, \quad i \in N, \tag{1}$$

где X^n – множество целых положительных чисел, размерность $n = |X^n|$, $x_i < +\infty$, N - множество натуральных чисел с размерностью $n = |N|$, $n < +\infty$, Подзадачей задачи (1) будем называть следующую задачу:

$$\sum_{i=1}^{k} \quad a_i x_i = S, a_i \in \{0,1\}, x_i \in X^n, a_i = 1, i \in K, a_i = 0, i \in N \setminus K, \tag{2}$$

где $X^k \subseteq X^n$, $k = |X^k|$, $k \leq n$, $K \subseteq N$, $k=|K|$, K – подмножество индексов всех выбранных переменных $x_i \in X^n$, $N \setminus K$ - подмножество индексов всех невыбранных переменных $x_i \in X^n$ подзадачи (2). Отметим, что задача (1) является частным случаем задачи (2), когда множество $K = \emptyset$.

Предложены полиномиальные алгоритмы решения задачи о сумме подмножеств (2) на основе работ [18, 19] с разделением исходного множества X^n на два подмножества. Затем применяется к этим подмножествам в отдельности алгоритм генерации сочетания C_n^m для получения необходимых подмножеств и потом используется метод слияния для полученных подмножеств, из которых находим искомое подмножество X^k. Здесь k=2m.

По результатам проведенного анализа существующих поисковых систем были сделаны выводы о том, что алгоритмы поиска частично удовлетворяют современным требованиям сбора, хранения, обработки и анализа больших объемов данных.

Все поисковые процессы должны протекать с соблюдением трех основных характеристик больших данных:

1. Volume
2. Velocity
3. Variety

Все три характеристики должны быть применены к базовой задаче, как правило, с несколькими переменными и нетривиальным условием. Работы с большими данными технологии: поддержка Hadoop, интеграция с программно-аппаратными комплексами.

Заключение

На основе полученных результатов можно сделать следующие выводы:

- приведенный обзор работ показывает, что важная задача о сумме подмножеств в теории сложности алгоритмов относится к основной из трудных проблем класса NP-complete и может быть использована по разработке поисковых систем;

- предложены полиномиальные алгоритмы решения задачи о сумме подмножеств (2) на основе работ [18, 19] с разделением исходного множества X^n на два подмножества;

- использованы алгоритм генерации сочетания C_n^m и метод слияния для получения искомого подмножества X^k.

ЛИТЕРАТУРА

1. Richard Bellman. Notes on the theory of dynamic programming iv - maximization over discrete sets. Naval Research Logistics Quarterly, 3(1-2):P.67–70, 1956.
2. David Pisinger. Linear time algorithms for knapsack problems with bounded weights. Journal of Algorithms, 33(1):1 – 14, 1999.
3. Konstantinos Koiliaris, Chao Xu. A Faster pseudopolynomial time algorithm for subset sum. To appear in SODA ’17, 2017. //arXiv:1610.04712v2[cs.Ds] 8 Jan 2017.-18p.
4. Karl Bringmann. A near-linear pseudopolynomial time algorithm for subset sum. To appear in SODA ’17, 2017. //arXiv:1610.04712v2[cs.Ds] 8 Jan 2017.-18p.
5. E. Horowitz, S. Sanni. Computing Partitions with Application to the Knapsack Problem //Journal of the ACM(JACM), 1974, T21, pp.277-292
6. R. Schroeppel, A. Shamir A $T=O(2^{n/2})$, $S=O(2^{n/4})$ Algorithm for Certain NP-Complete Problem // SIAM Journal on Computing, 1981, Vol.10, № 3, pp.456-464
7. Cobham A. The intrinsic computational difficulty of functions. //In Proceedinges of the Congress for logic, methodology and philosophy of science.-NorthHoiLand, 1964.P.24-30.
8. Egmonds J. Parths, treers and flowers. // Canadian Journal of mathematics. -1965. Vol.17. – P.449-467.
9. Николаев А.В. Геометрический подход к задаче о разрезе. -Ярославль, ЯрГУ, 2014, -68с.
10. Robert G. Jeroslow. Trivial integer programs unsolvable by branch-and-bound // Mathematical Programming. — 1974. — V. 6. — P. 105–109
11. Krishnamoorth B. Bounds on the size of branch-and-bound proofs for integer knapsacks // OR Letters. — 2008. — V. 36, № 1. — P. 19–25.
12. Колпаков Р.М., Посыпкин МА. Верхняя оценка числа ветвлений для задачи о сумме подмножеств. //Матем. Вопросы кибернетики. Вып.18.-М.: Физматлит, 2013.-с.213-226
13. C. J. van Rijsbergen. "Information Retrieval." Dept. of Computer Science. University of Glasgow, 1979
14. A. Adamansky. Overview of full-text search methods and algorithms. -Novosibirsk, Novosibirsk State University, 2018, -26p.
15. Min Chen, Shiwen Mao, Yin Zhang, Victor C.M. Leung. Big Data. Related Technologies, Challenges, and Future Prospects. — Spinger, 2014. — 100 p.
16. *Лифшиц Ю*рий. Точные алгоритмы и открытые проблемы // Современные задачи *теоретической информатики*. [Электронный ресурс] URL: http://download.yandex.ru/class/ (дата обращения: 15.01.2021)
17. Куликов Александр. Алгоритмы NP-трудных задач. Лекции. //Computer Science клуб Санкт-Петербургского отделения при Математическом Институте имени В.А. Стеклова, 2009. .-с.13-16
18. B. Sinchev, A.B. Sinchev, J. Akzhanova, A.M. Mukhanova. New methods of information search. I. // News of the National Academy of Sciences of Kazakhstan, Series of Geology and Technical Sciences, Volume 3, Number 435 (2019), pp. 240-246

19. B. Sinchev, A.B. Sinchev, J. Akzhanova, Y. Issekeshev, A.M. Mukhanova. Polynomial time algorithms for solving NP-complete problems . // News of the National Academy of Sciences of Kazakhstan, Series of Geology and Technical Sciences, Volume 3, Number 441 (2020), pp.97-101
20. S.G. Akl. Adaptive and optimal algorithms for enumerating perinutations and combinations. //The Computer Journal, 30 (1987), pp. 433-436
21. B. Sinchev, A. B. Sinchev, Z.A. Akzhanova. Computing network architecture for reducing a computing operation time and memory usage associated with determining, from a set of data elements, a subset of at least two data elements, associated with a target computing operation result. // Патент в USPTO, 2019.-38p.

Ауезова Ә.С. [1], Муратова К.Н. [1], Синчев Б. [1]

Құрылымданбаған деректерді ақпараттық іздеу әдістері

Аңдатпа. Мақалада құрылымданбаған (мәтіндік) деректерді ақпараттық іздеу мәселесін шешу үшін қолданылатын жаңа әдістер қарастырылған. Құжаттарды іздеу негізгі сөздер бойынша іздеу машиналарында қолданылатын табиғи тілде жүзеге асырылады. Осылайша көпмүшелік алгоритмдер негізінде уақыт пен кеңістік бойынша жақсартылған сипаттамалары бар бірнеше кілттер бойынша әмбебап іріктеу машинасы жасалады. Бұл ұсынылған машина экономиканың әртүрлі салаларында үлкен деректерді өңдеу үшін қолданылуы мүмкін. Мақсатқа жету үшін жаңа полиномиялық алгоритмдер негізінде NP-complete класына жататын ішкі жиындардың қосындысы туралы есеп пайдаланылды. Полиномиялық және экспоненциалды алгоритмдердің уақыты мен кеңістігінде, осы алгоритмдер жақсырақ және едәуір тиімді.

Түйінді сөздер: іздеу, әдіс, алгоритм, құрылымданбаған ақпарат, іздеу машинасы, көпмүшелік алгоритмдер, NP-толық, үлкен деректер (Big data).

Auyezova A.S. [1], Muratova K.N. [1], Sinchev B. [1]

Methods of information search for unstructured data

Abstract. The article discusses new methods used to solve the problem of information retrieval of unstructured (text) data. Search for documents is carried out by keywords in the natural language used in search engines. Thus, on the basis of polynomial algorithms, a universal multi-key sampling machine is created with improved time and space characteristics. This proposed machine can be used for processing big data in various areas of the economy. To achieve this goal, the new polynomial algorithms are based on the problem of the sum of subsets, which belongs to the NP-complete class. These algorithms are significantly more time and space efficient than the existing best polynomial and exponential algorithms.

Key words: search, method, algorithm, unstructured information, search engine, polynomial algorithms, NP-complete, big data.

Сведения об авторах:

Ауезова Анель Саттаровна, PhD-докторант, лектор кафедры «Информационных систем» Международного университета информационных технологий.

Муратова Камила Нурлановна, магистр, лектор кафедры «Информационных систем» Международного университета информационных технологий.

Синчев Бахтгерей, д.т.н., профессор кафедры «Информационных систем» Международного университета информационных технологий.

About authors:

Anel S. Auyezova, PhD-doctoral student, lecturer, «Information Systems» Department, International Information Technology University.

Kamila N. Muratova, Master of Tech. Sci., lecturer, «Information Systems» Department, International Information Technology University.

Bakhtgerey Sinchev, Doct. of Tech. Sci., Professor, «Information Systems» Department, International Information Technology University.

Авторлар туралы мәлімет:

Ауезова Әнел Саттарқызы, PhD-докторанты, Халықаралық ақпараттық технологиялар университетінің «Ақпараттық жүйелер» кафедрасының лекторы.

Муратова Камила Нурланқызы, магистр, Халықаралық ақпараттық технологиялар университетінің "Ақпараттық жүйелер» кафедрасының лекторы.

Синчев Бахтгерей, т.ғ.д., Халықаралық ақпараттық технологиялар университетінің «Ақпараттық жүйелер» кафедрасының профессоры.

УДК 530.1, 681.3.06

Нәдіров Н.Қ. [1], **Дуйсебекова К.С.** [1]

[1] Международный университет информационных технологий, Алматы, Казахстан

РАЗРАБОТКА СИСТЕМЫ ФОРМИРОВАНИЯ ПРОФИЛЯ КЛИЕНТА НА ОСНОВЕ BIGDATA С ИСПОЛЬЗОВАНИЕМ СЕМАНТИЧЕСКОГО АНАЛИЗА

Аннотация. *По мере того, как бизнес-процессы становятся все более и более ориентированными на клиента, чрезвычайно важны маркетинговые и управленческие отношения с клиентами, что заслуживает особого внимания с учетом глобализации и растущей конкуренции на рынке. На данном этапе возникает много всевозможных вопросов «Кто такие клиенты?» и «Как их идентифицировать?». В настоящее время «Какую информацию необходимо собирать о потенциальных клиентах?» и «Каковы сейчас уровни обслуживания?».*

Цель разработчика заглянуть в «голову» собственного клиента и осознать, чего он желает. Эти познания станут неоценимыми позднее, когда найденные данные и информация будут изучены, осознание всех задач и вопросов станет неотъемлемой частью успеха будущего проекта.

В данной статье представлен способ подготовительной обработки данных для умственного изучения данных, доктрина использования, способы изучения больших размеров данных и извлечения полезной информации, содержащейся в них полезной информации для формирования профиля клиента.

Ключевые слова*: Большие данные, программное обеспечение, информационная система, структурированные и неструктурированные данные, социальный профиль человека.*

Введение

Актуальность темы анализа состоит в том, что сейчас случается повсеместное проникновение IT-продуктов в жизнь современного человека независимо от его сферы работы. Все это дает большое число преимуществ, которые, непременно, положительно влияют на работу человека.

Обычно каждый проект с приложением проходит несколько шагов либо актуальных циклов от постановки трудности до её визуализации либо представления готового решения. В статье мы рассмотрим любой из этих шагов больше тщательно на примере усовершенствования продаж, принятная схема процесса представлена на Рисунке 1.