Optimization of Information Retrieval in Digital Libraries on the Basis of Influent Analysis

Marat Rakhmatullaev
Tashkent University of Information Technologies, Uzbekistan
marat56@mail.ru

Abstract

The aim of the research is to increase the efficiency of information retrieval for scientific and educational information in electronic libraries of corporate networks based on the application of the influent analysis method. Influential analysis allows: identify the most important indicators that need to be given the most attention, allocate funds to optimize the total efficiency score; Manage some factors to improve work of a system; Reduce the impact of those indicators that we can't influence (manage) at present time.The application of the method is especially effective for the development of a corporate network of academic libraries in order to improve the speed of information retrieval and the exchange of valuable scientific resources.

Keywords: digital library, influential analysis, automated library system, information support, data search optimization, databases, search criteria for information, information environment.

Ключевые слова. Электронная библиотека, инфлюентный анализ, автоматизированная библиотечная система, информационное обеспечение, оптимизация поиска данных, базы данных, критерии поиска информации, информационная среда.

Введение. Рост объема научно-образовательной информации приводит к проблеме их хранения и повышения оперативности поиска данных. Сужение области поиска наиболее важна в электронных библиотеках, т.к. требуется оперативность и достоверность данных для научных исследований, в образовательных процессах, а также в принятии технологических решений. Речь идет не только о локальных электронных базах данных, но и о корпоративных сетях информационно-ресурсных центров и библиотек.

Как известно, корпоративное взаимодействие существенно экономит средства на создание информационных ресурсов, исключая дублирование данных, позволяет делиться ценными знаниями, повысить оперативность поиска и др. Вместе с тем, повышается ответственность за накапливаемую информацию и ее распространение. Конечная цель любой информационной системы – максимальное удовлетворение пользователя, максимальное соответствие полученного результата его запросам. В статье рассмотрен подход к сужению объема информации на основе методов инфлюентного анализа и нечеткой логики для повышения эффективности поиска данных в электронных библиотеках.

Основная часть. В теории детерминированного инфлюентного анализа введено понятие инфлюенты – как оценки влияния изменения входных параметров на изменения выходных параметров в системах управления (Трухаев, 1984; Звягин, 2016). Хотя теория рассчитана в основном на производственные и экономические объекты, но основные принципы можно

использовать и в технологиях информационного поиска и оценки информации. При поиске информации различные показатели влияют на результаты по разному, «цена» в каждом случае неодинакова. Например, если вам нужны лишь научные достижения по той или иной области, наиболее существенным фактором является поиск в научной информационной среде, в научных журналах с высоким показателем импакт-фактора и др.

Но если вас интересуют лишь применение научного результата, например, в бизнес проектах, то область исследований будет уже другой. На практике бывает трудно определить, нашли вы искомую информацию или нет, т.е. насколько точно она соответствует не только вашим требованиям, но и требованиям времени. Поэтому использование как самих критериев поиска, так и их степени влияния очень важно при поиске искомой информации. Количественная оценка уровня влияния (инфлюентов) критериев очень трудно поддается формализации (Ibrahim, 2011; Рахматуллаев, 2012; Zhu, 2011).

Формирование информационных ресурсов происходит в некоторой информационной среде. Здесь под информационной средой (ИСд) понимается пространственно-временная область, содержащая средства и условия, необходимые и достаточные (технические, экономические, организационные и др.) для информационного обеспечения пользователей (главным образом, в научной и образовательной сферах деятельности) по определенным критериям (Рахматуллаев, 2012).

Основными характеристиками, которые отражают ИСд, являются:

1. Системы классификации и кодирования информации и стандарты (УДК, ББК, классификаторы, MARC форматы и т.д.)
2. Система формирования научных информационных ресурсов;
3. Система поиска информации;
4. Базы данных научной информации (электронный каталог, полнотекстовые базы данных научных журналов, книг и т.д.);
5. Система отображения информации (интерфейс, web pages и т.д.);
6. Система администрирования баз данных научной информации;
7. Разработчик информационных ресурсов (авторы, библиотекари, операторы, администраторы баз данных и т.д.);
8. Пользователь информационных ресурсов (юзеры);
9. Телекоммуникационная инфраструктура (локальные сети, Интранет, Интернет).

Корпоративная сеть электронных библиотек характеризуется следующими признаками: наличие нескольких источников информации (в нашем случае электронных библиотек); возможность дистанционного доступа; добровольность предоставления доступа к ресурсам; наличие программно-технической инфраструктуры для реализации активного информационного обмена. Она обладает различными параметрами, факторами, которые в разной мере могут влиять на конечный результат ее функционирования.

Рассмотрим наиболее важные из них:

1. Процедуры поиска и выдачи данных по запросу;

2. Объем баз данных научной информации;
3. Цена источника информации;
4. Обеспечение информационной безопасности;
5. Администрирование сети и баз данных;
6. Организационная структура;
7. Уровень квалификации кадров;
8. Техническая инфраструктура

В различных условиях информационной среды эти факторы по разному влияют на конечный показатель. Управление этими параметрами является важным аспектом достижения желаемого результата.

В инфлюентном анализе исследуются вопросы влияния тех или показателей на результаты поиска. Какие из характеристик ИСд в наибольшей степени влияют на качество поиска данных в базах данных, в локальных электронных библиотеках, в корпоративных информационно-библиотечных сетях?

Задача инфлюентного (факторного) анализа, формулируемая следующим образом (Трухаев, 1984): по заданной зависимости

$$Y = f(X) = F(X_1, X_2, \ldots X_n,) \quad (1)$$

показателя *Y* от факторов $X_1, X_2, \ldots X_n$, а также при известных начальных a_i и конечных b_i значениях факторов X_i определить величины A^f_x , являющиеся оценками влияния приращения $\delta_i = a_i - b_i$ фактора X_i на приращение

$$\Delta Y = \Delta f = f(b) - f(a) \quad (2)$$

результирующего показателя.

При этом величины A^f_x, называемые инфлюентами, должны удовлетворять условию

$$\Sigma A^f_x = \Delta f \quad (3)$$

В информационно-библиотечных системах факторы могут быть связаны с параметрами, характеризующими поиск информации, ее хранение, телекоммуникационную инфраструктуру, организационное обеспечение, элементы информационной безопасности, т.е. все то, что влияет на процесс нахождения нужных данных в электронных библиотеках.

В настоящее время имеются ряд математических методов, которые применяются для расчета инфлюент: Логарифмический метод, Метод деления нераспределенных остатков поровну; Метод обратных колец; Методы неопределенных множителей; Метод сопряженных множителей; Метод решающей матрицы ; Метод точки Лагранжа; Метод частных функций и т.д. Каждый метод имеет свои преимущества и недостатки в зависимости от условий их применения.

Для инфлюентного анализа в процессе поиска информации необходимо нахождение инфлюент A_x как оценок влияния изменений объема подобласти поиска на объем выводимых данных. При поиске информации необходимо выделить те критерии, которые в той или иной степени отражают поисковые образы и «отсекают» информационные слои, которые не соответствуют запросу или соответствуют несущественно.

Выделим эти критерии:

R - Поисковый образ (ключевые слова, автор и др.);

P - Профессия пользователя (его интересы, область исследований и др.)

K - Уровень квалификации (школьник, студент (бакалавр или магистрант), докторант, преподаватель, доцент, профессор и др.);

S - Социальный статус (безработный, работающий и т.д.)

T- Цель поиска (бизнес, научные исследования, преподавание и др.)

S - Социальный статус (безработный, работающий и т.д.)

T- Цель поиска (бизнес, научные исследования, преподавание и др.)

Т.о. мы имеем функцию Y, отражающую конечную цель, результат поиска данных в базе данных БД:

$$Y = f(R, P, K, S, T)$$

Эти критерии в той или иной степени влияют на процесс нахождения источника информации. Формирование запросов происходит на основе применения оценок с использованием методов нечетких множеств, экспертного опроса и инфлюентного анализа, которые позволяют выделить наиболее важные показатели для поиска и «отсечь» ненужные области данных. Они могут сыграть существенную роль при сокращении области поиска информации в электронных библиотеках.

Инфлюентный анализ и защита информационных ресурсов. При хранении информационных ресурсов нас интересует не столько объем информации, а его важность, соответствующая запросу пользователей. Соответственно, уместно определить его значимость, цену. Это имеет важное значение не только при поиске данных, но и в создании систем защиты информационных ресурсов.

Для оценки ресурса выделим наиболее важные критерии A_i с указанием четырех показателей оценки M_J:

A1 - начальная стоимость ресурса (0 – бесплатный; 1- ниже среднего; 2- средняя; 3- выше среднего);

A2 - ресурсная среда (0 – отдельно на магнитных носителях или на компьютере; 1 - локальная сеть, 2 - корпоративная сеть, 3 - глобальная сеть);

A3 - тип ресурса (0 - общедоступная информация, 1 - корпоративная информация, 2 - платная информация, 3 - конфиденциальная информация);

A4 - использование ресурсов (0- очень редко; 1 редко; 2 – часто; 3 – очень часто);

A5 - важность ресурса во всей информационной системе (0 - сбой ресурса (его разрушение) не влияет на другие ресурсы; 1 - сбой ресурса приводит к остановке некоторых компонентов других ресурсов; 2 - сбой ресурса приводит к сбою нескольких компонентов других ресурсов; 3 - сбой ресурса приводит к сбоям в работе всей системы;

A6 – затраты на восстановление ресурса (0 - незначительные, 1 - средние, 2 - высокие, 3 – большие затраты или невосстанавливаемые);

A7 - время, затрачиваемое на восстановление ресурса (0 - незначительное, 1 - среднее, 2 – требует много времени, 4 – требуется критически много времени);

А8 - возможность частично или полностью восстановить ресурс (0 - легко, 1 - требует времени и средств; 2 - требует значительных временных и материальных затрат; 3 – невозможно восстановить);

А9 - нарушение конфиденциальности (0 - не наносит существенного ущерба; 1 - наносит материальный ущерб в определенных обстоятельствах; 2 - приводит к меньшему моральному и / или материальному ущербу; 3 - причиняет значительный ущерб или отказ всей системы);

А10 - нарушение целостности данных ресурса (0 - не приводит серьезным последствиям; 1 – результаты заметны (ощутимы), но не приводят к прекращению работы; 2 - Результатом будет неисправность; 3 - следствием этого является искажение данных ресурса и невозможность их исправить).

При оценке ресурсов привлекаются эксперты высокой квалификации для установки условной цены для каждого показателя M_j. В качестве примера в Таблице 1 показана оценка значимости каждого показателя по 5-ти бальной шкале (от 0 до 4) . Соответственно, 0 – самая низкая оценка , 4 - самая высокая.

Табл. 1

Условные цены информационных ресурсов

Критерии	Показатели оценки ресурса			
	M_1	M_2	M_3	M_4
A_1	0	0-1	1-2	2-4
A_2	0	0-1	1-2	2-4
A_3	0-1	0-2	1-3	3-4
A_4	0	0-1	1-2	2-4
A_5	0	0-1	1-2	2-4
A_6	0-1	1-2	1-3	2-4
A_7	0	1-2	2-3	2-4
A_8	0-1	0-2	1-3	3-4
A_9	0-2	1-2	2-3	2-4
A_{10}	0-1	0-2	1-3	3-4
$\sum M_i$	**0-6**	**4-16**	**12-26**	**23-40**

Определяя для каждого ресурса суммарные показатели, мы можем судить о цене (значимости) рассматриваемого источника информации и принимать меры как по хранению, так и по его защите.

Влияние фактора (инфлюента) «категория пользователя» на процесс поиска информации также имеет место. Кто делает запрос? Каковы его интересы? профессия? уровень образования? и т.д. Разным категориям пользователей, разных профессий и уровней при одном и том же запросе нужная разная информация, результат.

Результаты опросов, проведенный усилиями сотрудников Ташкентского университета информационных технологий для 567 респондентов различных областей деятельности показывают уровень интересов по различным категориям пользователей от видов литературы (их интересов) по 100 бальной шкале. Для категории пользователей Научный сотрудник, Преподаватель, Инженеры, Соискатели ученых степеней, магистранты и докторанты , Студент, Бизнесмен был проведен опрос по интересам Научные исследования, Диссертации, Энциклопедии, Учебники, учебно-методические материалы, Материалы по проф. Технологиям, Энциклопедии, Статьи по технологиям, Статьи по бизнесу, Стандарты.

Табл.2

Уровень интересов читателей различных областей деятельности

№	Категория интересов и материалы	Пользователь					
		Научный сотрудник	Преподаватель	Инженеры	Соискатели ученых степен магистранты докторанты	Студент	Бизнесмен
		1.	2.	3.	4.	5.	6.
1.	Научные исследования(статьи и книги)	98.3	82.5	40.2	99.4	20.1	12.6
2.	Статьи по технологиям	22.4	32.1	99.2	70.3	56.6	42.2
3.	Диссертации,	99.6	80	56.7	98.7	70.6	10.4
4.	Материалы по проф. технологиям	51.4	50	99.6	50.4	47.3	21.3
5.	Энциклопедии	92.4	63.9	51.3	54.3	42.4	11.2
6.	Стандарты	10.6	32.5	90.4	41.5	21.5	52.4
7.	Учебники, учебно-методические материалы	71.4	100	43.1	72.8	99.7	12.5
8.	Статьи по бизнесу	22.5	20	47.1	31.6	13.1	99.8

Соответственно, при получении запросов мы можем учитывать разные категории пользователей и их интересы. Этот критерий может существенно снизить объем рассматриваемой при поиске информации.

Пусть Q – объем информации в ИС.

$V = \{v_1, v_2\ v_3 \ldots v_n\}$

v_i – это категория интересов пользователя. Например, научные исследования, диссертации, учебная литература и др.

$K = \{ k_1, k_2, k_3, \ldots k_m, \}$

k_i – категория пользователей (научный сотрудник, преподаватель, студент и др.).

Конечно, разделение по кругу интересов – это не догма и потребность в той или иной информации может меняться и расширяться. Но в целом можно выявить те или иные

приоритеты при поиске информации, сузив области и ускорить нахождение искомых данных.
Инфлюентный анализ для выявления наиболее значимых показателей для поиска информации в корпоративных библиотечных сетях включает следующие этапы:

1. Систематизация и классификация информации в базах данных корпоративной информационно-библиотечной сети;
2. Проведение экспертного опроса на предмет выявления наиболее существенных факторов, влияющих на поиск данных.
3. Инфлюентный анализ факторов. Выявление параметров, которые в той или иной мере влияют на процесс формирования и поиска данных;
4. Внесение изменений в процедуры (алгоритмы и программы поиска) поиска данных с учетом инфлюентных факторов;
5. Feedback. Т.е. возможность итерации при систематизации и оценке ресурсов для повышения эффективности поиска.

Прикладные аспекты. Практическая реализация инфлюентного анализа в сочетании с методами нечеткой логики дает возможность повысить эффективность в поиске информации в электронных библиотеках, особенно в корпоративных сетях. Корпоративная сеть академических библиотек – информационно-ресурсных центров вузов Узбекистана для совместного использования электронных библиотек формируется на основе информационно-библиотечной системы ARMAT++ и Интранет телекоммуникационной сети Государственной программы «Электронное образование».

Система существенно повысила эффективность реализации сети и использует технологические решения, основанные на Cloud Technology, инфлюентного анализа и нечеткой логики. Корпоративная сеть ИРЦ объединяет более 60 вузов, поэтому очень важно использование инфлюентного анализа для оценки и отбора наиболее важных доступа параметров к наиболее ценной информации. Использование Cloud Computing в сочетании с методами инфлюентного анализа в информационно-библиотечных сетях может существенно повысить эффективность как обслуживания пользователей, так и администрирование сети.

Эксперименты, проделанные в корпоративной библиотечной сети вузов Узбекистана с использованием системы ARMAT++ показали существенные преимущества применения таких технологий: высокая скорость обработки библиотечных данных; снижение технических требований к персональным компьютерам в ИРЦ и пользователей; отказоустойчивость; экономичность и др.

При создании корпоративной библиотечной сети вузов формируется информационное пространство на головном сервере. В нашем случае он расположен в Центре внедрения электронного образования (ЦВЭООУ при МинВУЗ). Информационное пространство предоставляет возможность хранить как Главную базу данных корпоративных информационных ресурсов, так и персональные данные библиотек – членов консорциума.

Администрирование базы данных происходит централизованно. Пополнение базы данных новыми ресурсами производится при поддержке Министерства среднего и специального образования (МВССО) РУз, которое выделяет финансовые средства на сканирование

учебников, учебно-методических пособий и электронную каталогизацию. Сводный электронный каталог формируется на основе MARC21 коммуникативного формата. Библиотеки, работающие на разных коммуникативных форматах, могут использовать конвертор и стандарт ISO 2709.

Инфлюентный анализ может быть применен не только при формировании источников информации (баз данных, электронных библиотек и др.), но и при исследовании степени использования информационных ресурсов, причин снижения или повышения публикационной активности в организациях и даже в масштабах страны. Нахождение различных факторов и их численных значений могут производиться на основе опроса респондентов в той или иной области. Например, в Узбекистане ежегодно производится опрос ученых, докторантов, преподавателей и исследователей на предмет изучения проблем использования электронных научно-образовательных ресурсов и публикационной активности.

Как показали социологические исследования о влиянии тех или иных факторов на публикационную активность в Узбекистане, наибольшее влияние оказывает:

1. 28% - Сам стимул написания статьи для научных сотрудников, преподавателей, докторантов. Учет публикаций в высокорейтинговых журналах при аттестации, влияние на карьерный рост;

2. 27% - Наличие ценного научного материала, научный потенциал. Уровень научных исследований в организации, наличие лабораторий с современным оборудованием и др.

3. 25% - Навыки написания научных статей. Умение оформить статью под формат тех престижных журналов, под которые подходит тема статьи.

4. 15% - Знание иностранных языков. Главным образом – английского языка, который является языком общения научных сообществ.

5. 5% - Информированность о научных журналах. В настоящее время этот показатель все меньше и меньше влияет на публикационную активность, т.к. списки научных журналов можно легко найти в Интернете, в списках Web of Science и Scopus.

При исследовании влияния факторов на использование научных журналов с высоким рейтингом в научных исследовниях наиболее важным показателем оказались требования к диссертационным работам и защитам для соискателей ученых степеней. ВАК установил требования по научным публикациям: не менее 3-х научных статей в журналах из списка Web of Science и Scopus. Кроме того, при аттестации преподавателей и научных работников учитываются публикации в престижных научных журналах. За этот показатель проголосовали 30% опрошенных.

Повышение уровня доступа к научным журналам ведущих мировых издательств – 25%. За последние годы Национальная библиотека Узбекистана и другие ведущие министерства и ведомства выделяют достаточно весомые финансовые средства на подписку на электронные базы научных журналов таких издательств и агрегаторов, как EBSCO Information Services (национальная подписка), Springer, Wiley, ProQuest (для некоторых вузов, библиотек) и др. Но имеется необходимость в расширении подписки на некоторые издания, имеющие важное значение для ученых и преподавателей вузов и научных центров. С 2018 г. имеется подписка на ресурсы компании Elsevier для всех вузов республики (как на аналитическую систему Scopus, так и научные журналы).

Знание английского языка (более 20%) остается актуальным для докторантов и молодых ученых при изучении, анализе научных публикаций, соответственно и написании научных статей для престижных журналов, а также для подготовки грамотных презентаций. Но с каждым годом этот показатель снижается, т.к. все больше вузов и научных центров создают группы обучения на английском языке, а также привлекают преподавателей из зарубежных стран для чтения лекций, что также влияет на данный критерий.

Повышение интереса к научным исследованиям (10%) остается пока низким стимулирующим фактором (зарплата, условия для ученых и т.д.). При этом, имеются большие проблемы в создании условий для научных исследований (научные лаборатории, оборудование, финансирование поездок на научные конференции и др.), которые требуют существенных финансовых вложений.

Востребованность в высококвалифицированных кадрах, которые постоянно работают над собой, имеют навыки проведения научных исследований, хотя и медленно, но растет (15%). Это связано с тем, что в республике активно развивается промышленность, внедряются наукоемкие технологии, особенно государство уделяет большое внимание на развитие информационных технологий и телекоммуникаций. Создаются технопарки, открываются новые научные центры. Соответственно растет востребованность в кадрах с высоким уровнем знаний.

Как показывает опыт ведущих стран именно развитие наукоемких производств и создание соответствующих рабочих мест являются важными факторами получения эффективного конечного результата.

Заключение.

Исследования проводятся на базе Ташкентского университета информационных технологий для разработки новых программных комплексов информационно-библиотечной системы ARMAT+ для корпоративной информационной сети библиотек вузов Узбекистана (Rakhmatullaev, 2010). Использование методов инфлюентного анализа позволяет существенно сократить объем поиска за счет отсекания массивов данных неадекватных запросу. Недостаток метода – сложность получения объективных оценок ресурсов от экспертов; усложнение программного комплекса по формированию и поиску информации; наличие большого числа субъективных параметров; не всегда методы инфлюентного анализа дают рациональное решение, иногда приходится перепроверять результаты расчетов.

Реализация новых дополнительных функций по оценке ресурсов очевидно сопряжена с дополнительными затратами, с потребностью повышения квалификации персонала, с дополнительными напряжениями всех участников информационного обмена, особенно в начальный период работы корпоративной сети.

Инфлюетный анализ позволяет:

- выявить наиболее важные показатели, на которые нужно уделять наибольшее внимание, выделять средства для оптимизации суммарного показателя эффективности;
- управлять некоторыми факторами, чтобы улучшить работу всей системы;
- снизить влияние тех показателей, на которые мы в данный момент не можем повлиять (управлять);

- пренебречь теми показателями, которые несущественно влияют на процесс поиска и нахождения искомой информации.

Список литературы

Звягин, Л. С. (2016). Методы инфлюентного анализа и принятие решений, Технические науки: проблемы и перспективы: материалы IV Междунар. науч. конф. (г. Санкт-Петербург, июль 2016 г.) = [Influential analysis methods and decision making, Technical sciences: problems and prospects]. — СПб: Свое издательство. — vi, 134 с. ISBN 978-5-4386-0975-9.

Рахматуллаев, М.А. (2012). Информационная среда инновационных образовательных технологий в вузах = [Information environment of innovative educational technologies in universities]. Международный научно-практический семинар “Активизация деятельности университетов в создании инновационной среды в национальной экономике”. Фергана. ФГУ. 2012. 205-207.

Трухаев, Р.И. (1984). Инфлюэнтный анализ и принятие решений (детерминированный анализ) = [Influential analysis and decision making].— М.: Наука.

Ibrahim, J. G., Zhu, H.T., Tang, N. S. (2011). Bayesian local influence for survival models (with discussion). Lifetime Data Analysis, 17, 43-70.

Rakhmatullaev, M. (2010). Development of information technologies in libraries of Uzbekistan. Reforms and their results. Bibiliotechyi Visnyk. Scientific theoretical applied journal. N2. National Library of Ukraine. 22-25.

Zhu, HT., Ibrahim JG, Tang NS. (2011). Bayesian influence approach: a geometric approach. Biometrika, 98, 307-323.

Zhu, H.T., Ibrahim, J.G., Lee, S.Y., and Zhang, H.P. (2007). Appropriate perturbation and influence measures in local influence. Annals of Statistics, 35, 2565-2588.