

ПРОБЛЕМА КАЖУЩЕЙСЯ ОЧЕВИДНОСТИ В ОНЛАЙНОВОМ ИНФОРМАЦИОННОМ ПОИСКЕ

А. А. Рагойша

Белорусский государственный университет

ragoisha@bsu.by

A PROBLEM OF ILLUSIVE EVIDENCE IN ONLINE INFORMATION RETRIEVAL

A. Rahoisha

Previous experience may encourage students to follow ineffective strategy when they proceed from general to specialized information retrieval; a common problem is their belief that Google is the only suitable tool to deal with any kind of search tasks. A special training should be undertaken to persuade the students to perform actions that seem obvious to the scientist. The issues to be discussed in class include uncertain quality of retrieved documents, simplifying procedures of evaluation of information, shortcomings of ranking based on the popularity of the resource, the need for critical analysis of data that is published in edited sources.

ЗНАНИЕ: Google — это:

- огромная база данных,
- эффективный алгоритм ранжирования,
- безупречная техническая поддержка (высокая скорость работы, бесперебойность).

**ВЕРА:
Google
ЗНАЕТ
ВСЕ**

ЗНАНИЕ: Интернет анархичен.

Google — зеркало Интернета (не кривое!).

Google легко находит.

Человек тщательно перепроверяет найденное.

Пример. *Кто автор фразы:*

"Широко простирает химия руки свои в дела человеческие"

241 документ — за Ломоносова.

Google "широко простирает химия" ломоносов -мендел Search [Advanced Search](#)

Web [Show options...](#) Results 1 - 10 of about 241 for "широко простирает химия" ломоносов -менделеев.

[manaev: "Широко простирает химия" руки свои в дела - \[Translate this page \]](#)
"Широко простирает химия" руки свои в дела человеческие". - Ломоносов. (Post a new comment). [info] lee_van_cleef 2006-07-11 07:08 pm UTC (link) ...
manaev.livejournal.com/634362.html

342 документа — за Менделеева?!.

Google "широко простирает химия" менделеев -ломонс Search [Advanced Search](#)

Web [Show options...](#) Results 1 - 10 of about 342 for "широко простирает химия" менделеев -ломоносов.

[Химия и жизнь \(Шимшон Шиманский\) / философия / Проза.ру ... - \[Translate this page \]](#)
6 фев 2010 ... "Широко простирает химия" руки свои в дела человеческие".
(Д.И.Менделеев) Есть большая группа элементов в Периодической таблице Менделеева, ...
www.proza.ru/2010/02/06/1145 - [Cached](#)

**Правда
крупно
проиграла**

Результаты поиска улучшаются при тщательном рассмотрении:

- (1) *n*-ная доля из обнаруженных выше 342 документов все же *не* приписывает Д.И.Менделееву авторства, например:

[\[PDF\] Лекция № 1 - \[Translate this page \]](#)
File Format: Microsoft Powerpoint - [View as HTML](#)
Широко простирает химия руки свои в дела человеческие...» Выделение газа, изменение окраски, Согласно уравнению Клайперона – **Менделеева ...**

но не обошлось без микродезинформации

- (2) Если запрос состоит только из фразы "широко простирает химия...", в начале извлеченного списка свидетельств в пользу Ломоносова больше, чем в пользу Менделеева.
Здесь заслуга Google очевидна.

Критерий оценки достоверности документа: **КТО АВТОР?**

Учитывать — ДА, бездумно верить — НЕТ.

Пример — две соседние записи в списке результатов:

На любительском автомобильном сайте — правда

[Денисов Алексей Аполлинарьевич. Автомобили, зима и химия](#)
"Широко простирает химия руки свои в дела человеческие" - говаривал Михайло Ломоносов. Эх, знал бы он, насколько! Значение химии особенно ощущают ...
[zhurnal.lib.ru/.../avtomobilizima.shtml](#) - [Сохранено в кэше](#) - [Похожие](#)

[ХИМИЧЕСКИЙ ФАКУЛЬТЕТ - КАФЕДРЫ - Кафедра общей и неорганической химии](#)
... необъятным возможностям неорганической химии, наглядно показывает насколько "широко простирает химия руки свои в дела человеческие" (Д. И. Менделеев). ...
[www.chimfak.rsu.ru/KAF/neorg/neorg.htm](#) - [Сохранено в кэше](#)

На профессиональном сайте химфака — неправда

Google в библиографическом поиске

Задача. Известно полное библиографическое описание статьи.

Найти первоисточник.



Что же в списке результатов поиска ?

Первоисточник отсутствует.

Обнаруженные документы всего лишь цитируют эту статью.

Переформулировать запрос?

Перебор вариантов → Затраты усилий с неясной перспективой.

Как действовать?

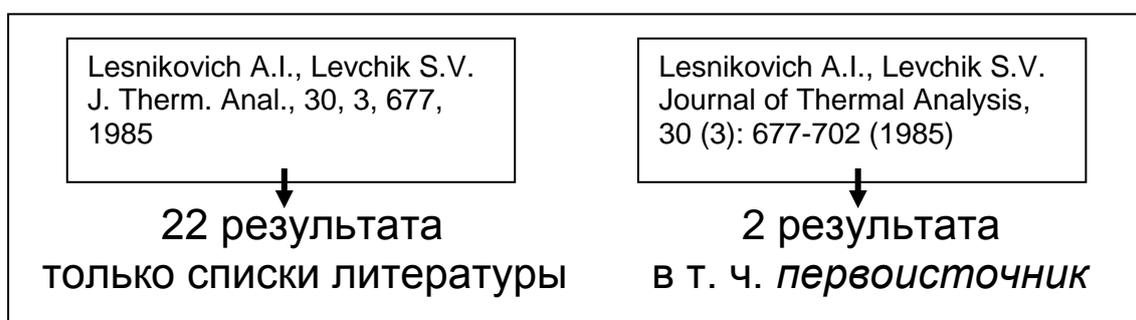
1. *Осознать*: не мы первые попали в Интернет.
Задолго до нас профессионалы разрабатывали вспомогательные поисковые инструменты и алгоритмы поиска, отбирали материал для специализированных баз данных, формировали тематические метасайты, рецензировали научные работы.
2. *Не изобретать велосипед*. Не пытаться начинать с нуля решение любой поисковой задачи. Максимально использовать наработки профессионалов, по крайней мере, на стадии первичного отбора материала.
Одно из следствий: стремиться использовать (только) рецензируемые информационные источники.
3. Сэкономленные умственные усилия направить на **анализ качества и содержания** извлеченных документов.

Пример: библиографический поиск.

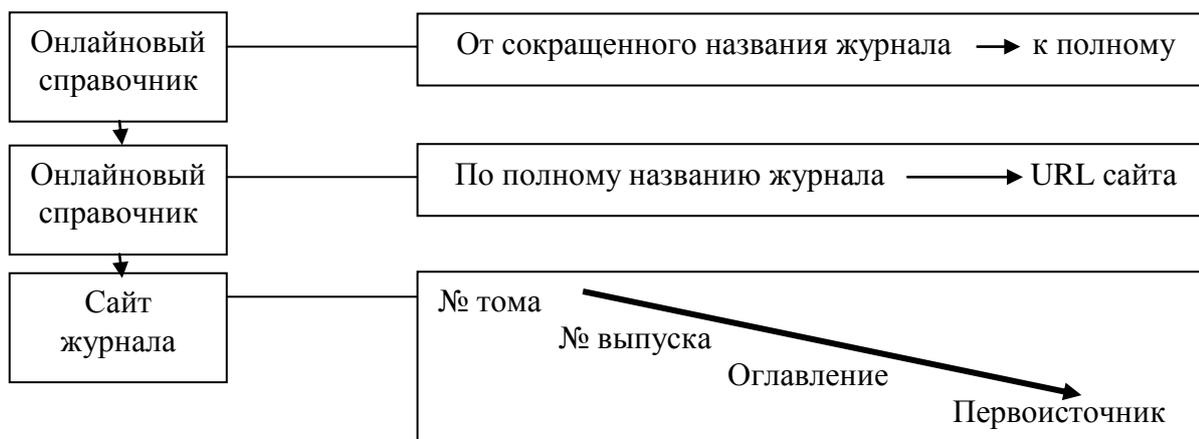
Поиск первоисточника по известному библиографическому описанию следовало бы проводить в библиографической базе данных — полнотекстовой либо содержащей URL (DOI) статей.

Эффективные инструменты (доступность ограничена):
Scopus, ISI Web of Knowledge.

Удовлетворительный доступный инструмент: Google Scholar.



Универсальный альтернативный алгоритм:
в центре внимания — сайт журнала.



В рутинной работе алгоритм практически всегда приводит к успеху, но **нелюбим студентами** за его многостадийность.

Достоверность и авторитетность автора (издателя)

Бесплатных хороших ресурсов — все больше.

Но ... беды роста.

- повальный переход от осмысленного ручного труда к автоматизации;
- гонка за количеством методом *Copy&Paste*;
- пренебрежительное отношение к редакторской и корректорской работе;
- безответственность:
принцип "as is" — общепринятая норма.

Грань между категориями "редактируемая база данных" и "репозиторий" размывается. Размываются границы между достоверным, сомнительным и недостоверным.

Иллюстрации из очень хороших баз данных.

Пример 1. Физические свойства H_2SO_4 в *ChemSpider* (экспериментальные данные):

The image shows a screenshot of the ChemSpider website for H_2SO_4 . The page title is H_2SO_4 . Under the heading "Experimental Physchem Properties", there is a list of properties:

- ⊕ Melting Point: close to 0 C (depends upon concentration) ?
- ⊕ Melting Point: -2 C ?
- ⊕ Boiling Point: 554F ?
- ⊕ Boiling Point: close to 100 C (depends on concentration) ?
- ⊕ Boiling Point: 327 C ?
- Freezing Point: 51F ?
- Specific Gravity: 1.84 ?

Annotations with arrows point to specific data points:

- A box on the left contains the text: "Плавится при $-2^{\circ}C$
Замерзает при $+10^{\circ}C$ ". Arrows point from this box to the "Melting Point: -2 C" and "Freezing Point: 51F" entries.
- A box on the right contains the text: "Кипит при $290^{\circ}C$
Кипит при $327^{\circ}C$ ". Arrows point from this box to the "Boiling Point: 554F" and "Boiling Point: 327 C" entries.
- A box at the bottom right contains the text: "Тонкий юмор в справочнике?". An arrow points from this box to the "Boiling Point: 327 C" entry.

ChemSpider позиционирует себя как ведущий общемировой центр в области структурной органической химии.

С 2009 г. — это собственность *Royal Society of Chemistry*. Авторитетность *RSC* не вызывает сомнений, и вполне закономерно, что доверие пользователя автоматически распространяется на весь *ChemSpider*. Как показывает практика, такой подход не всегда оправдан.

В приведенном выше примере пользователь должен *догадаться*, что перед ним не рекомендованные составителем, а механически собранные из Интернета числа. Проблема оценки их достоверности *перекладывается* на самого пользователя.

Пример 2. Артефакты, в WWW превращающиеся в факты

В первоисточнике могут встречаться обозначения, условность которых оговаривается в поясняющих документах.

После копирования в иные базы данных эти элементы приобретают для непосвященного пользователя черты реальных параметров.

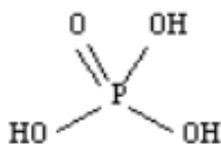
Растворимость уксусной кислоты в воде:
1000 г/л
(в исходной базе данных так обозначается
смешиваемость во всех отношениях)

```
CAS Number : 000064-19-7
Chem Name   : ACETIC ACID
Mol Formula : C2H4O2
Mol Weight  : 60.05
Melting Pt  : 16.6 deg C
Boiling Pt  : 117.9 deg C
Water Solubility:
  Value     : 1E+006 mg/L
  Temp      : 25 deg C
  Type      : EXP
```

Chemical Structure:

Registry Number: 7601-54-9

Formula: $\text{H}_3\text{O}_4\text{P} \cdot 3\text{Na}$



• 3 Na

Структурная формула
трехзамещенного фосфата натрия
(в старых базах данных CAS так было
принято отображать структуру солей)

Достоверность и онлайнное рецензирование

Надежды, что коллективный контроль качества онлайнной информации окажется эффективным, за редким исключением, не оправдались.

При оценке достоверности документа результат коллективной экспертизы следует считать не решающим, а совещательным.

Пример
(ChemSpider).

Каков
CAS
Registry
Number
этана?

Names and Synonyms

Database ID(s)

Validated by Experts, Validated by Users, Non-Validated,
Approved by Experts

270-652-0 [EINECS/ELINCS]

271-259-7 [EINECS/ELINCS]

68475-58-1 [RN]

68527-16-2 [RN]

эксперты неправы

Ethanato

Ethane

Ethane [UN1035] [Flammable gas]

Ethane, refrigerated liquid [UN1961] [Flammable gas]

1730716 [Beilstein]

200-814-8 [EINECS/ELINCS]

270-651-5 [EINECS/ELINCS]

271-734-9 [EINECS/ELINCS]

68475-57-0 [RN]

68606-25-7 [RN]

7261-70-3 [RN]

74-84-0 [RN]

Правильный
ответ
находится
в конце
списка

Журнал, практикующий онлайнное рецензирование статей и считающийся рецензируемым, на практике может содержать *нерецензированные публикации*.

e-LC Documents for 2009

- **Properties of non-symmetric bent-core liquid crystals with variable flexible chain length**
2009/Nov/24 17:15:25
[Katalin Fodor-Csorba](#), Michal Kohout, Martin Chambers, Aniko Vajda, Galli, Attila Domján, Jiří Svoboda, Alexey Bubnov, Antal Jáklj.
[Abstract \(54 hits\)](#) [Full Document \(144 hits\) \(405 Kbytes\)](#) [Comments](#)

Пример.

Число
комментариев: **0**