

Онтологическое аннотирование библиографических ссылок в научных публикациях

М.Р. Когаловский

E-mail: kogalov@cemi.rssi.ru

Институт проблем рынка РАН

Москва

Работа поддержана грантом РГНФ , проект 11-02-12026-в

Конференция «Проблемы наукометрии. Состояние и перспективы развития»,
Москва, 10-12 октября 2013 г.

Вместо введения

- Распространение открытого доступа к публикациям результатов исследования в онлайн-режиме привело к использованию ряда новых метрик в наукометрических исследованиях, оценивающих, например, востребованность публикаций (загрузки и просмотры), количество созданных «закладок» и др.
- Несмотря на это, библиографические ссылки, традиционно используемые в качестве основы разнообразных метрик в этой области, остаются главным инструментом в наукометрии.
- Потенциал ссылок цитирования в наукометрии не использован в полной мере. Пути расширения возможностей его использования обсуждаются в докладе.
- Эта работа является составной частью проекта «Соционет»

Ссылки цитирования в научных публикациях

- Ссылки на используемые источники в научных публикациях (ссылки цитирования) - норма научной этики, ее нарушение трактуется как плагиат
- Коллекции ссылок цитирования, вместе с тем, служат источником информации для наукометрических измерений
- Индексы цитирования публикаций, индексы Хирша, импакт-факторы журналов рассматриваются как оценки влияния публикаций и научных сотрудников на развитие науки, как критерии авторитетности изданий
- В научном сообществе в последнее время складывается критическое отношение к этим оценкам и критериям, осознается необходимость использования их с осторожностью
- Примеры - *The San Francisco Declaration on Research Assessment - DORA, подходы к оценке научной продуктивности в некоторых европейских странах*
- Тем не менее, индексы цитирования продолжают широко использоваться для оценки научной результативности отдельных ученых и исследовательских организаций.

Конференция «Проблемы наукометрии. Состояние и перспективы развития»,
Москва, 10-12 октября 2013 г.

«Немые» ссылки цитирования

- Сложившаяся практика использования ссылок цитирования в наукометрии имеет, по крайней мере, два «слабых места»
- Первое связано с тем, что в качестве источников данных для оценки индексов цитирования и для других наукометрических исследований служат ссылки из пристатейных списков использованных источников
- Но библиографические ссылки в пристатейных списках - «немые», они не несут какой-либо информации о мотивах цитирования, их информационный потенциал ограничен
- Некоторые парадоксальные последствия их использования
- Второе слабое место – не учитывается, что *на один и тот же источник в тексте может существовать несколько ссылок с разными мотивами цитирования*
- Пристатейные списки «схлопывают» эту множественность
- Как преодолеть эти «слабые места»?

Семантические ссылки цитирования vs «немые» ссылки

- Альтернатива «немым» ссылкам - семантические ссылки
- **Семантическая ссылка** – ссылка с определенной явным образом семантикой, характеризующей мотивы цитирования
- Семантические ссылки обеспечивают новые возможности для наукометрических исследований :
 - утонченная статистика,
 - результаты анализа топологии графа семантических ссылок
- Вместе с тем, сохраняется возможность оценки на их основе традиционных индексов цитирования.

Определение семантики ссылок цитирования

- В каких терминах можно определить семантику ссылки?
- Для этой цели можно использовать онтологии ссылок цитирования CiTO (Shotton D.), SWAN и SKOS (W3C) и другие, разработанные в последние годы, и аннотировать ссылки в их терминах
- Простейшая форма онтологии ссылок – таксономия ссылок (иерархический классификатор)
- Таксономию ссылок можно представить в виде набора контролируемых словарей классов ссылок
- Такие словари реализованы в системе Соционет (не только ссылки цитирования, но и другие виды семантических связей)
- Для спецификации семантики ссылки можно использовать автоматизированный анализ эмоционального тона контекста ссылки в тексте публикации (для оценочных ссылок) и экспертный подход.

Пример словаря классов ссылок

1. Словарь оценочных ссылок

1.1.	Позитивно оценивает цитируемый источник
1.2.	Негативно оценивает цитируемый источник
1.3.	Нейтрально оценивает цитируемый источник
1.4.	Оценивает как новаторскую работу
1.5.	Оценивает как интересный результат
1.6.	Оценивает как заблуждение

Семаническое аннотирование ссылок

- **Аннотирование** – ассоциирование с данным информационным ресурсом метаданных, характеризующих его свойства (*например, теги в Википедии – категоризация статей*)
- **Семантическое аннотирование** – это аннотирование *семантическими метаданными*, т.е. метаданными, которые описывают семантику ресурса
- Частный случай семантического аннотирования - **онтологическое аннотирование**
- Как конструктивно осуществить аннотирование ссылок?
- Онтологическое аннотирование ссылок в пристатейном списке источников **некорректно**: *на один и тот же источник в тексте может существовать несколько ссылок с разной семантикой*
- Поэтому **объектами онтологического аннотирования** должны быть *ссылки или представители ссылок внутри текста*, они определяют семантические связи публикаций.

Разновидности библиографических ссылок

- Разновидности ссылок в отечественных публикациях определяет ГОСТ 7.0.5-2008 «Система стандартов по информации, библиотечному и издательскому делу. Библиографическая ссылка. Общие требования и правила составления». Издание официальное. М.: Стандартинформ, 2009.
- Стандарт предусматривает такие разновидности ссылок:
 - внутритекстовые библиографические ссылки
 - подстрочные библиографические ссылки
 - затекстовые библиографические ссылки
- Представление библиографических ссылок в тексте:
 - внутритекстовые ссылки - непосредственно
 - подстрочные ссылки - знаки сноски
 - затекстовые ссылки - знаки отсылки и выноски.

Конференция «Проблемы наукометрии. Состояние и перспективы развития»,
Москва, 10-12 октября 2013 г.

Примеры ВИДОВ ССЫЛОК: *внутритекстовая ссылка*

WordNet is perhaps the most popular thesaurus in use today. WordNet is a lexical database for the English language, with over 42,000 terms. It can be used interactively online (URL: <http://www.cogsci.princeton.edu/~wn/>). WordNet was developed by cognition experts, adopting psycholinguistic theories of human lexical memory.

Примеры
ВИДОВ ССЫЛОК:
*подстрочная
ссылка*

В перспективе для сокращения бюджетного дефицита РФ **предлагается⁷** снизить расходы бюджета в реальном выражении, поскольку расширение государственных заимствований ведет к росту процентных ставок и вытеснению инвестиций частного сектора. Но «одна лишь констатация возможной опасности чрезмерных заимствований

⁷ Основные направления финансовой политики на 2011 г. и период 2012-2013 гг. - С.7.

**Представители
подстрочных
ссылок в тексте:
*знак сноски***

В перспективе для сокращения бюджетного дефицита РФ **предлагается⁷** снизить расходы бюджета в реальном выражении, поскольку расширение государственных заимствований ведет к росту процентных ставок и вытеснению инвестиций частного сектора. Но «одна лишь констатация возможной опасности чрезмерных заимствований

Примеры ВИДОВ ССЫЛОК: затекстовая ссылка — 1

The monitoring is performed by means of sensors that measure parameters of the patient state: pulse rate, blood pressure, and temperature. The distributed monitoring system should not only ensure registration and safe storage of the data but also selectively deliver them to the doctor who takes care of the patient, as well as to control emergence of dangerous situations.

The character of problems of this kind can be illustrated by the example of systems for monitoring the health state of patients who lead an active life but need to be under permanent control of medical personnel [84]. The data stream integration technique gives us a regular way of creation of multilevel distributed media for delivery of information, making thus easier development of various applications based on sensor networks. Sensor networks are classified among one of the directions of the distributed computing, Pervasive Computing, which is closely related to the information grid [26].

6. CONCLUSIONS

The direction of distributed computing—grid—considered in the paper solves the problem of development of the technique for work of applications with a set of distributed resources. The approach is based on the concept that the resources are integrated in an infrastructure (grid) in which the access to resources is uni-

ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research (project nos. 06-07-89111 and 08-07-00381), by the Presidium of Russian Academy of Sciences (the basic research program), and by a Russian Federation Presidential Grant from the Program for Support of Leading Scientific Schools (project no. NSh-2139.2008.9).

REFERENCES

1. Open Grid Forum Organizational Members. http://www.ogf.org/Members/members_members.php.
2. Stockinger, H., Defining the Grid: A Snapshot on the Current View, *Swiss Institute of Bioinformatics (Vital-IT)*, CH-1015 Lausanne, Switzerland, 2006.
3. Stockinger, H., Grid Computing: A Critical Discussion on Business Applicability, *IEEE Distributed Systems Online*, 2006, vol. 7, no. 6.
4. Schikuta, E. et al., Business in the Grid: Project Results, *The 1st Austrian Grid Symp.*, 2005. www.pri.univie.ac.at/Publications/2005/Schikuta_austrian-grid_bigresults.pdf.
5. *Grids—a Low Cost, High Speed Alternative to Traditional High-Performance Computing*, 2003, Grid Technology Partners.

Примеры ВИДОВ ССЫЛОК: затекстовая ссылка – 2

cas on. The most important of these are: emphasizing precision over recall, focusing on search results instead of documents, combining the strengths of local and global indices, applying clustering, and using relevance feedback.

While peer-to-peer technology has seen widespread adoption in file sharing, it is not often used for solving information retrieval problems. This is unfortunate as it has the potential to offer a robust solution to the ethical and technical problems that plague centralized solutions and thus deserves more attention. We believe one of the primary reasons for its unpopularity in science is because the field lacks a clear definition that distinguishes it from related and overlapping fields like general peer-to-peer systems and federated information retrieval. Moreover, there was no overview of what has been done so far and what choices can be made when implementing a practical system. Both of these aspects are important for future research and development of real-world systems. This article provides a solid basis for this.

REFERENCES

- ABERER, K. AND HAUSWIRTH, M. 2002. An overview on peer-to-peer information systems. In *Proceedings of the Workshop on Distributed Data and Structures (WDAS'02)*.
- ADAMIC, L. A., LUKOSE, R. M., PUNIYANI, A. R., AND HUBERMAN, B. A. 2001. Search in power-law networks. *Phys. Rev. E* 64, 4.
- AKAVIPAT, R., WU, L.-S., MENCZER, F., AND MAGUITMAN, A. G. 2006. Emerging semantic communities in peer Web search. In *Proceedings of the Workshop on Information Retrieval in Peer-to-Peer Networks (P2PIR'06)*. ACM Press, New York. 1–8.
- BALKE, W.-T., NEJDL, W., SIBERSKI, W., AND THADEN, U. 2005. Progressive distributed top-k retrieval in peer-to-peer networks. In *Proceedings of the International Conference on Data Engineering (ICDE'05)*. IEEE Computer Society, Los Alamitos, CA, 174–185.
- BAWA, M., MANKU, G. S., AND RAGHAVAN, P. 2003. Sets: Search enhanced by topic segmentation. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, New York, 306–313.

ACM Transactions on Information Systems, Vol. 30, No. 2, Article 9, Publication date: May 2012.

Конференция «Проблемы наукометрии. Состояние и перспективы развития»,
Москва, 10-12 октября 2013 г.

**Представители
затекстовых
ссылок в тексте:
*знаки отсылки - 1***

недостаточно. Некоторые сведения можно почерпнуть из [2–5]. Авторы учли также свой опыт создания морфолого-лексического анализатора [6]. При-

**Представители
затекстовых
ссылок в тексте:
*знак отсылки - 2***

ционно-коммуникационные технологии. Особое внимание было обращено на необходимость обеспечить широкий доступ к услугам и информации всех слоев населения, для чего необходимо постоянно учитывать политические, социально-экономические и культурные контексты, в которых реализуются стратегии перехода к электронному правительству [22, с. 56].

**Представители
затекстовых
ссылок в тексте:
*знаки отсылки -3***

complement their lacks such as Object Oriented, Multimedia, Deductive, Spatial models, etc. Simultaneously, knowledge level modeling techniques ([Uschold 1998](#)) propose alternative models to define information richer in expressiveness but also in complexity ([Wand et al. 1999](#); [Borgida 1995](#)). Consequently these representations applied mostly in AI, like

Объекты онтологического аннотирования

- Итак, объектами онтологического аннотирования должны быть:
 - индивидуальные или множественные:
 - внутритекстовые библиографические ссылки
 - знаки сноски
 - знаки отсылки
 - знаки выноски
- Множественные ссылки обычно имеют единую семантику
- Семантику внутритекстовых ссылок и ссылок других видов, представляемых в тексте знаками сноски, отсылки и выноски, можно чаще всего извлечь из контекста ссылки или указанного знака.

Возможный формат аннотации ССЫЛКИ

- (Recker, M. (2006). Perspectives on Teachers as Digital Library Users. *D-Lib Magazine*, September 2006) **#1.3**
- В перспективе для сокращения бюджетного дефицита РФ предлагается⁷**#1.2** снизить расходы бюджета в реальном выражении, поскольку расширение государственных.....
- [14 **#1.3**]
- [12, стр.5 **#1.1**]
- (Smith 2013 **#1.5**)
- [22 **#1.4**, 25 **#1.4**] или [22, 25] **#1.4**
- (Wand et al. 1999 **#1.2**; Borgida 1995 **#1.2**) или (Wand et al. 1999; Borgida 1995) **#1.2**
- Инструмент для аннотирования ссылок в интерактивном режиме может быть более дружелюбным и скрывать от пользователя такие «заклинания».

Конференция «Проблемы наукометрии. Состояние и перспективы развития»,
Москва, 10-12 октября 2013 г.

Итоги

- Семантика библиографических ссылок в настоящее время в наукометрии не учитывается
- Учет семантики ссылок позволяет получать существенно более содержательные наукометрические результаты
- В частности, наряду с новыми статистическими данными также представляют интерес и результаты анализа топологии графа семантических связей
- Репозитории аннотированных библиографических ссылок из текстов научных публикаций могут стать качественно новыми источниками информации для наукометрии
- Такой подход реализован в системе Соционет
- Не учитывается также множества ссылок на один и тот же источник, различным образом мотивированных
- Целесообразно разработать и ввести в деятельность научного сообщества стандарты таксономий библиографических ссылок и форматов представления их онтологических аннотаций.

Конец

Спасибо за внимание

Конференция «Проблемы наукометрии. Состояние и перспективы развития»,
Москва, 10-12 октября 2013 г.