ZESZYTY NAUKOWE UNIWERSYTETU SZCZECIŃSKIEGO

NR 789 STUDIA INFORMATICA NR 32 2013

Jakub Swacha[*]

Szczecin University

A SIMPLE TAXONOMY FOR COMPUTER SCIENCE PAPER RELATIONSHIPS

Summary

The high and increasing number of research papers published every year makes it very difficult to follow the current state of the knowledge in any given branch of science. Querying research databases on relevant keywords often returns thousands of results, whereas using too specialized keywords may result in omitting important papers. In this paper we propose a simple taxonomy of relationships between research papers and show how it can be used to improve retrieval of relevant papers, providing examples illustrating the potential benefits from its usage. We also discuss possible implementation scenarios and required software functionalities. Although the taxonomy is designed for papers from the area of computer science, it can be adapted for other branches of science.

Keywords: research paper relationship, publication relationship taxonomy, computer science publications, research database

Introduction

The scientists have at least four good reasons to publish research papers: dissemination, registration, validation, and designation of their scientific work (cf. Clarke 2010). As a result, the large scientific community produces a vast volume of publications: almost 30,000 active research journals (Tenopir et al. 2011) publish over one million papers yearly (Björk et al. 2009), and this amount does not include book chapters or conference proceedings: Institute of Electrical and Electronics Engineers alone admits to publish over 1000 proceedings every year (IEEE 2012). And the long term tendency is growing (Guptaet al. 1995).

[*] jakubs@wneiz.pl

The mere volume of publication makes it very difficult to follow the current state of the knowledge for any given topic. The readers need some kind of precise filtering mechanism and the best they can get is querying research publication databases on relevant keywords. The problem is they often get thousands of matching results that could not be further filtered out manually unless spending enormous amount of work, whereas using too highly specialized keywords may lead to omitting important sources.

A traditional way of looking for papers related to a given topic is to:

- find a non-empty initial set of relevant papers, and then
- check their references, and
- look for papers that reference them.

Most of the modern research publication databases see e.g., ACM Digital Library (Hennessey 2012), CiteSeerX (Li et al. 2006), Google Scholar (Hoseth 2011), or IEEE Xplore (O'Neill 2010) support all stages of this procedure.

The initial set of relevant papers can be found with either search and indexing services provided with the database or web search engines (Brophy, Bawden 2005) using precise keywords, so that the results can be checked manually with a high chance of finding a truly relevant paper. Even if a web search engine was used in the first step, the procedure has to be continued within research publication database environment that lists both the references of a given paper and publications which cite that paper. Often a list of "related papers" is also available, which may also be considered, though often it contains too few or far too many items.

As a result of this procedure one gets a list that consists of tens rather than thousands of potentially relevant papers. Still, the mere fact of reference link between a relevant paper and another one does not guarantee the latter is also relevant. Many of papers from the list would not be considered for reading had the user known the actual kind of relationship between the two papers.

1. Contribution

We propose a simple taxonomy that defines the types of relationship between two research papers. Not only does it let the user instantly identify the kind of relationship and select actually relevant papers to follow, but it also gives a general idea of how the specific topic developed in the literature, distinguishing both theoretical and practical contributions.

The proposed taxonomy takes into account only semantic and not formal types of relationships, such as shared keywords, same author, same journal, etc. This is because the relevance of the paper to a given topic does not depend on formal relationships unless they are backed by semantic relationships. The formal types of relationships are already modeled in research publication databases, hence including them in the proposed taxonomy would increase the complexity of the taxonomy without providing any additional benefit.

For the same reason the taxonomy deals only with the relationships between papers, not the papers themselves. We are aware of the variety of properties the papers may have and subclasses the papers may belong to (see e.g. D'Arcus, Giasson 2009 or Table 3 by Shotton 2010). Yet the proposed taxonomy is in no way an attempt to replace existing taxonomies and metadata specifications; on the contrary, it is envisaged as an information retrieval aid expected to coexist with them, and simplicity is its key value.

As the focus is on the relationships between papers (considered as single entities), the taxonomy also does not cover the internal structure or types of content elements within the papers.

The taxonomy has been conceived for papers in the field of computer science, yet it could be used as-is for most of the applied sciences. It would require some level of adaptation in order to become useful for other branches of science.

The key contributions of this paper are:

- specification of requirements for an taxonomy of relationships between research papers that would make it useful in retrieving computer science papers,
- definition of the taxonomy of relationships between research papers in the field of computer science, with accordance to the specified requirements,
- providing relationship examples to illustrate the benefits of having related papers arranged according to the proposed taxonomy,
- discussion of implementation scenarios and software functionalities required in order to make use of the proposed taxonomy.

2. Related work

In their classic work Moravcsik and Murugesan (1975) define four types of citations: a) *conceptual* or *operational*; b) *organic* or *perfunctory*; c) *evolutionary* or *juxtapositional*, and d) *confirmatory* or *negational*. In the same year, Chubin and Moitra (1975) published their typology which consists of six types of citations: a) *affirmative essential basic* (the referenced paper is declared central to the reported research); b) *affirmative essential subsidiary* (referencing a method or tool that is not directly connected to the subject of the paper, but is still essential to the reported research); c) *affirmative supplementary additional* (referencing supportive observation with which the citer agrees); d) *affirmative supplementary perfunctory* (related to the reported research without additional comment); e) *negational partial* (a citer suggests that the referenced paper is erroneous in part and offers a correction); f) *negational total* (a citer refers to the referenced paper as being completely wrong).

Hidetsugu Nanba et al. investigate the relationships between scientific papers in their work on automatic generation of a review (or survey) articles in a specific subject domain (Nanba et al. 2000). They distinguish three types of citations: *B* (those that show other researchers' theories or methods for the theoretical basis), *C* (those that point out the problems or gaps in related works) and *O* (other).

Similar categories are proposed by Teufel (2001) in her research on the evaluation of automatically generated summaries. She defines seven types of sentence's "rhetorical status", of which three correspond to references to other papers (*contrast*, *basis*, *other*), with the former two being of primary importance to the reader.

Bertin et al. (2006) in their research on qualitative evaluation of research works propose five categories of citations: a) *point of view* (asserting opinion); b) *comparison* (with subcategories: *resemblance* and *disparity*); c) *information* (with subcategories: *hypothesis*, *analysis*, *result*, *method*, *citation*, *counter-example*); d) *definition*, and e) *appreciation* (*positive* or *negative*).

Jörg (2008) extends the typology of Moravcsik and Murugesan with the following classes: *foundational* (inspired by), *state-of-the-art* (currently most works focused on), and *experimental* (explore).

Shotton (2010) proposes Citation Typing Ontology that can be used for annotation of reference lists and visualization of citation networks. He distinguishes 23 kinds of relationship between citing and cited document, which are arranged into two groups: *factual* and *rhetorical*. The former include: *cites, cites as authority, cites as metadata document, cites as source document, cites for information, is cited by, obtains background from, shares authors with, uses data from, uses method in*; the latter are arranged into three subgroups, covering respectively positive (*confirms, credits, extends, obtains support from, supports, updates*), negative (*corrects, critiques, disagrees with, qualifies, refutes*), and neutral (*discusses, reviews*) relationships.

Recently, Newman, Bechhofer and De Roure (2009) described an ontology for myExperiment, a "Social Virtual Research Environment" capable of facilitating management and sharing of Research Objects, supporting a social model, providing an open extensible environment and a platform to action research. Because of these aims, it is very complex, and completely different than the taxonomy proposed in this paper.

3. Paper relationship taxonomy

3.1. Specification of requirements

Considering the overall aim of developing the taxonomy, that is aiding in the process of retrieving relevant papers in the field of computer science, as well as trying to keep the taxonomy as simple as possible, the following requirements were proposed:

1. *Focus on the relationships between papers*. The taxonomy should not cover anything but the relationships between papers (for instance, it should not model the papers themselves).
2. *Focus on the semantic relationships*. The taxonomy should not cover formal relationships between papers (such as, e.g., reference, shared keywords, same author, same journal, etc.).
3. *Treat a publication as a whole*. The taxonomy should not consider the structure of individual papers or investigate the relationships between its elements.
4. *Include only concepts which are useful for the assumed application* (i.e., filtering related computer science papers based on their relevance). For instance, Shotton's *shares authors with* relationship is not one.
5. *Distinguish the classes of relationships that usually make difference for the user*. For instance, the three types of relationships defined by Nanba are too general to be useful in the information retrieval context.
6. *Do not define too specific classes of relationships* – so that the user is able to grasp the entire taxonomy. The 23 classes proposed by Shotton seem to be too many.
7. *Let the classification of relationships be unambiguous.* The user should never have doubt what kind of relationship there is between two given

papers, or what to expect from a paper in a specific kind of relationship with a given one.

8. *Let the classification of relationships be objective.* For this reason, the relationships should not be distinguished based on author's stance (positive or negative) which in some contexts may be hard to identify.
9. *Make the labels of relationships classes self-describing.* For instance, Moravcsik and Murugesan's conceptual or operational are not.
10. *Make the labels of the relationships classes easy to comprehend.* For instance, Chubin and Moitra's affirmative supplementary perfunctory is not.

3.2. Specification of the taxonomy

After carefully analyzing the specified requirements, a taxonomy of the relationships between computer science papers has been designed consisting of one root class (*paper relationship*) having two obligatory properties: *subject* (*A*) and *object* (*B*). These properties should unambiguously identify (in the context of specific application) the two papers in the relationship. We assume no constraints on their type or value except for that their values must be set.

The paper relationship class has ten subclasses: applies, extends, generalizes, implements, improves, mentions, quotes, specializes, tests, and theoretizes:

1. The *applies* subclass denotes a relationship between papers *A* and *B*, in which paper *A* applies to a new domain a solution (method, algorithm, system) described in paper *B*.
2. The *extends* subclass denotes a relationship between papers *A* and *B*, in which paper *A* describes a solution that adds new functionalities to a solution (method, algorithm, system) described in paper *B*.
3. The *generalizes* subclass denotes a relationship between papers *A* and *B*, in which paper *A* adapts, to a more general domain, a solution (method, algorithm, system) described in paper *B*.
4. The *implements* subclass denotes a relationship between papers *A* and *B*, in which paper *A* implements, in a working system, a solution (method, algorithm, system) described in paper *B*.
5. The *improves* subclass denotes a relationship between papers *A* and *B*, in which paper *A* improves quality and/or performance of a solution (method, algorithm, system) described in paper *B*.
6. The *mentions* subclass denotes a relationship between papers *A* and *B*, in which paper *A* mentions in the text or lists as a reference paper *B*. In other

words, this subclass is for all relationships based on a direct reference without actual content citation that cannot be classified into any of the other classes.

7. The *quotes* subclass denotes a relationship between papers *A* and *B*, in which paper *A* cites text, results or figures from paper *B*. In other words, this subclass is for all relationships based on a direct reference with actual content citation that cannot be classified into any of the other classes.
8. The *specializes* subclass denotes a relationship between papers *A* and *B*, in which paper *A* adapts, to a more specific domain, a solution (method, algorithm, system) described in paper *B*.
9. The *tests* subclass denotes a relationship between papers *A* and *B*, in which paper *A* empirically examines, verifies or validates a solution (method, algorithm, system) described in paper *B*.
10. The *theoretizes* subclass denotes a relationship between papers *A* and *B*, in which paper *A* investigates theoretical background for a solution (method, algorithm, system) described in paper *B*.

There may be more than one relation defined for any two papers *A* and *B*. For instance, paper *A* may at the same time extend and improve the solution described in paper *B*, or test and theoretize on it; it can even describe both its generalization and further specialization.

Notice that apart of the *quotes* and *mentions* subclasses, all the remaining types of relationships between papers are modeled upon the relationships between the ideas the papers convey. They are semantic relationships that often are but need not be backed by references.

Regarding the *quotes* and *mentions* subclasses, they were conceived as means of failsafe classification of relationships whose existence is backed by references, but which either does not belong to any of the defined subclasses, or, during relationship database development, is not (yet) established by the classifying person.

3.3. Examples of paper relationships

In this section we shall present exemplary relationships between papers to illustrate better the individual subclasses of relationships defined in the proposed taxonomy. We shall base our example on papers relevant to a classic work in computer science, *Data Compression Using Adaptive Coding and Partial String Matching* by J.G. Cleary and I.H. Witten (all the papers mentioned in this section will be referred to using only their title and author names for the sake of brevity). At the time of writing these words, Google Scholar listed 832 citations for this paper; twenty of them were chosen to illustrate each type of relationship with two examples (see Table 1).

After looking at Table 1 it should be easy to imagine how helpful arranging the papers within the proposed taxonomy would be for a user looking for papers relevant to a given topic. For instance, a user interested in state of the art should focus mainly on *improves* relationships; a user looking for adaptations to a specific sub-domain should focus mainly on *specializes* relationships; and a user interested in theoretical background should focus mainly on *theoretizes* relationships.

Table 1

Real-world examples of paper relationships

Subclass	Examples in relation to J.G. Cleary and I.H. Witten's *Data Compression Using Adaptive Coding and Partial String Matching*
Applies	*A PPM-like, tag-based branch predictor* by P. Michaud *Web Prefetching Using Partial Match Prediction* by T. Palpanas
Extends	*Constructing Word-Based Text Compression Algorithms* by N. Horspool and G. Cormack *Unbounded length contexts for PPM* by J.G. Cleary and W.J. Teahan
Generalizes	*An Executable Taxonomy of On-Line Modeling Algorithms* by S. Bunton *On prediction using variable order Markov models* by R. Begleiter et al.
Implements	*Implementing the PPM data compression scheme* by A. Moffat *The Design and Analysis of Efficient Lossless Data Compression Systems* by P.G. Howard
Improves	*PPM: one step to practicality* by D. Shkarin *Semantically Motivated Improvements for PPM Variants* by S. Bunton
Mentions	*Constructing Suffix Arrays of Large Texts* by K. Sadakane and H. Imai *Extracting key-substring-group features for text classification* by D. Zhang
Quotes	*Modeling for text compression* by T. Bell *Sequential weighting algorithms for multialphabet sources* by Tj.J. Tjalkens et al.
Specializes	*Compressing XML with multiplexed hierarchical PPM models* by J. Cheney *PPMexe: PPM for Compressing Software* by M. Drinić and D. Kirovski
Tests	*Experiments on the Zero Frequency Problem* by J.G. Cleary and W.J. Teahan *State of the art concerning Lossless Medical Image Coding* by K. Denecker et al.
Theoretizes	*Compression, Information Theory and Grammars: A Unified Approach* by A. Bookstein and Sh.T. Klein *Relationship Between Hidden Markov Models And Prediction By Partial Matching Models* by S.A. Yeates

Source: own elaboration.

4. Towards practical application of the taxonomy

4.1. Concepts for implementation

The definition of the relationship classes that was presented in the previous section is merely the first step in providing the benefits of the taxonomy to the users. It can be compared to a skeleton, whereas the flesh should consist of actual relationships between published papers. Obviously, identifying and classifying the relationships requires considerable effort.

There are basically three scenarios for implementation of the proposed taxonomy in the environment of a research publication database:

1. Making it a core component of a newly built system.
2. Embedding it as a new functionality of an existing system.
3. Providing a new, alternative front-end for an existing system.

Ad. 1. This approach has dual benefits. From a technical viewpoint, it would allow seamless integration of the proposed taxonomy in the system without need for any adaptation or cooperation. From an economic viewpoint, it would allow to use financial and/or work resources arranged for system development on acquiring the paper relationships information. This approach however requires a system developer willing to incorporate the taxonomy in the newly built system.

Ad. 2. The advantage here is that the benefits from the taxonomy would become immediately available for many users (of the existing system). These many users would also allow a cost-free acquisition of the paper relationships information – provided the users could classify the references they themselves followed. Technically, it would require adaptation of the existing system to integrate the new module. This approach requires a publication database administrator willing to add the new functionality into the existing system.

Ad. 3. In this scenario, a new front-end system would be developed, consisting of a user interface, a paper relationship database and relevant algorithms. The paper-related data would be fetched from an existing system. There are open access publication databases that make this approach feasible without any cooperation with developers or administrators of the existing system. This approach requires financial and/or work resources for acquiring the paper relationships information, because it would be unrealistic here to rely on the community effort as in scenario #2, as the key feature that could attract new users would only become useful after substantial amount of the paper relationships information was gathered.

4.2. Software implementation

The software implementing paper retrieval aided by the proposed taxonomy should have the following functionalities:

- allowing the user to specify queries with filtering or grouping based on paper relationships,
- allowing the user to view results of queries with (possibly interactive) filtering or grouping based on paper relationships,
- allowing the user to edit types of relationships between papers.

With the ongoing progress in natural language processing technologies, an optional software module could be developed, capable of an automatic or semi-automatic (i.e., formulating suggestions to be confirmed by a user) identification and classification of the relationships between papers.

Conclusions

In this paper we have defined a simple taxonomy of relationships between research papers. It allows for filtering related papers based on the nature of their relationships. In real-world information retrieval scenarios, it may lead to considerable decrease of time spent by users on acquiring set of research papers relevant to a given topic.

The proposed solution optimizes the way user interacts with research database system rather than the algorithms used to retrieve the documents. Still, it requires a software component to be implemented.

There are three practical scenarios for its implementation, each with advantages and disadvantages of its own.

Acknowledgements. The author wishes to thank Szymon Grabowski for his proofreading and both useful and funny comments.

References

Bertin M., Desclés J.-P., Djioua B., Krushkov Y., *Automatic annotation in text for bibliometrics use*. In: *Proceedings of the 19th International Florida Artificial Intelligence Research Society Conference*, AAAI Press, Melbourne Beach 2006.

Björk B.-C., Roos A., Lauri M., Scientific journal publishing: yearly volume and open access availability. *Inf Res* 2009, No. 14 (1), p. 391. http://InformationR.net/ir/14-1/paper391.html [accessed on 21.04.2012].

Brophy J., Bawden D., Is Google enough? Comparison of an internet search engine with academic library resources, *Aslib Proc* 2005, No. 57 (6) pp. 498–512, DOI: 10.1108 /00012530510634235.

Chubin D.E., Moitra S.D., Content analysis of references: adjunct or alternative to citation counting, *Soc Stud Sci* 1975, No. 5 (4), pp. 423–441, DOI: 10.2307/i212562.

Clarke M., Why hasn't scientific publishing been disrupted already? *The Scholarly Kitchen* 2010, http://scholarlykitchen.sspnet.org/2010/01/04/why-hasnt-scientific-publishing-been-disrupted-already [accessed on 21.04.2012].

D'Arcus B., Giasson F., Bibliographic Ontology Specification. Specification document, 4 November 2009, http://bibliontology.com/specification [accessed on 21.04.2012]

Gupta B., Sharma L., Karisiddappa C., Modelling the growth of papers in a scientific specialty, *Scientometrics* 1995, No. 33 (2), pp. 187-201, DOI: 10.1007/BF02020568.

Hennessey C.L., ACM Digital Library, *The Charleston Advisor* 2012, No. 13 (4), pp. 34–38, DOI: 10.5260/chara.13.4.34.

Hoseth A., Google Scholar, *The Charleston Advisor* 2011, No. 12 (3), pp. 36–39, DOI: 10.5260/chara.12.3.36.

IEEE Conference Proceedings (2012) http://www.ieee.org/conferences_events/conferences/xplore_conference_proceedings.html [accessed on 21.04.2012].

Jörg B. *Towards the nature of citations*. In: *Formal Ontology in Information Systems. Fifth International Conference. Poster Proceedings*, DFKI, Saarbrücken 2008.

Li H., Councill I., Lee W.-C., Giles C.L., *CiteSeerx: an architecture and web service design for an academic document search engine*. In: *Proceedings of the 15th international conference on World Wide Web*, ACM, New York 2006, DOI: 10.1145/1135777.1135926.

Moravcsik M.J., Murugesan P., Some results on the function and quality of citations, *Soc Stud Sci* 1975, No. 5 (1), pp. 86–92, DOI: 10.2307/i212559.

Nanba H., Kando N., Okumura M., *Classification of Research Papers using Citation Links and Citation Types: Towards Automatic Review Article Generation*. In: *Proceedings of the 11th SIG Classification Research Workshop*, ASIS, Chicago 2000.

Newman D.R., Bechhofer S., De Roure D., *myExperiment: An ontology for e-Research*. In: *Proceedings of the Workshop on Semantic Web Applications in Scientific Discourse*, SWSA, Washington 2009.

O'Neill A., IEEE Xplore 3.0: Is it new to you?, *IEEE Solid-State Circuits Mag* 2010, No. 2 (2), p. 6, DOI: 10.1109/MSSC.2010.936566.

Shotton D., CiTO, the Citation Typing Ontology, *J Biomed Semant* 2010, No. 1 (Suppl 1), p. S6, DOI: 10.1186/2041-1480-1-S1-S6.

Tenopir C., Mays R., Wu L., Journal article growth and reading patterns, *New Rev Inf Netw* 2011, No. 16 (1), pp. 4–22, DOI: 10.1080/13614576.2011.566796.

Teufel S., *Task-based evaluation of summary quality: Describing relationships between scientific papers*. In: *NAACL Workshop on Automatic Summarization*, ACL, Pittsburgh 2001.

PROSTA TAKSONOMIA ZWIĄZKÓW MIĘDZY PUBLIKACJAMI Z OBSZARU INFORMATYKI

Streszczenie

Wraz z wysoką i stale rosnącą liczbą publikacji naukowych publikowanych każdego roku, coraz trudniejsze staje się śledzenie aktualnego stanu wiedzy z wybranego obszaru badań. Przeszukiwanie baz publikacji naukowych, mimo odpowiedniego dobrania słów kluczowych, często zwraca tysiące rezultatów, których ręczne filtrowanie wymaga ogromnego nakładu pracy, a użycie zbyt precyzyjnych słów kluczowych może prowadzić do pominięcia ważnych źródeł. W artykule zaproponowano prostą taksonomię związków między publikacjami z obszaru informatyki i wyjaśniono, jak może być ona wykorzystana w celu poprawy wyszukiwania publikacji naukowych na wskazany temat, pokazując przykłady ilustrujące potencjalne korzyści z jej wykorzystania. Ponadto poddano dyskusji możliwe scenariusze implementacji i wymagany zakres funkcjonalności oprogramowania. Chociaż przedmiotem niniejszych rozważań są publikacje z informatyki, zaproponowane rozwiązanie może być z łatwością zaadaptowane do innych gałęzi nauki.

Słowa kluczowe: związki między publikacjami naukowymi, taksonomia związków pomiędzy publikacjami, bazy publikacji informatycznych