Algorithms to Detect Monolingual and Cross-lingual Similarity and Plagiarism

Ph.D. Theses

Máté Pataki

Supervisor:
Gábor Prószéky, D.Sc.

Doctoral School of Multidisciplinary Engineering Sciences,
Faculty of Information Technology,
Pázmány Péter Catholic University

Firenze, 2011

Budapest, 2012

Introduction

Plagiarism causes serious problems not only in higher education but also in a number of other professional fields. As the computer-administered papers are spreading and the students become acquainted with the computer and internet at a younger and younger age, plagiarism is leaking into secondary schools as well. Unfortunately, plagiarized articles and thoughts can more and more frequently be found in the academic world too. The spread of digital libraries are delayed by the illegal copies also because the authors – not entirely without reason – are afraid of the loss of their income. Also publishers often insist on paper-based publications because in this case it is much easier to limit illegal copying. The website contents of companies or even whole websites are more and more often reproduced by competing firms. Even Wikipedia, the largest online encyclopedia is struggling with plagiarism. The materials uploaded to Wikipedia are available for everyone free of charge and anyone can contribute to it by uploading contents. This is why Wikipedia administrators have to regularly monitor the content because they must evade anyone (even with good intent) uploading some unauthorized, copyrighted contents to their site.

Plagiarism detection today cannot be done without the help of computers. No one can know all the published works, articles, theses, websites on the given subject. In the case of a thesis it is not enough to feel that the work is plagiarized but it must be proved too. This requires a tool which is able to scan a huge amount of material in a short time and can name the sources used for the given thesis and the degree of matches.

Technical solutions providing protection against plagiarism can basically be divided into two groups, tools preventing copying (copy protection), and tools making the detection of copies possible (plagiarism detection). It is difficult to protect digital content from illegal copying without making the legal use more complicated in the meantime too. Moreover, in some cases, it is difficult even to provide access for everyone to the content regardless of the software environment used by them. Most of the copy protection systems are easy to pass by so they provide only nominal protection; other systems provide a better protection, it is complicated to pass them by but can only be used together with complementary softwares and in certain cases with dedicated hardware, which the user will only install or buy if the protected content is really valuable for them. People with disabilities (blind, visually impaired, deaf, the ones using old and outdated computers ...) are often not able to reach these protected contents, so in certain cases these procedures may even be infringing (Act no. XXVI of 1998, Section 6).

Plagiarism detection does not protect the content from illegal copying but if it is widely used, the path of the work is traceable and can prevent anyone referring to it as their own work. This protection is dual: on the one hand, finding a copy the system will immediately name the original source and the degree of the overlap; on the other hand, if the existence of such a system is widely known and is frequently used, most people will not take the risk of plagiarism since they do not want to expose themselves to the risk of being caught.

Methodology

There are several types of plagiarism detection systems and most of them can be used in certain fields efficiently, however, certain restrictions apply to most of them due to which they cannot be used in case of digital libraries or university theses, for instance. During my research I took two basic guidelines into consideration. The procedures and algorithms which I am developing and creating should make it possible to process large amounts of text, and similarity and plagiarism detection should be carried out with them automatically without human intervention. Another important aspect was language-independence and to make the algorithm operate for the Hungarian language. The latter is also important because at the beginning of the research such systems did not exist and were not available in Hungary at all.

In each case I tried to build in the algorithms resulting from my research into an operating and testable system, this way demonstrating not only their functionality but also promoting the domestic spread of the subject.

New scientific results

Plagiarism detecting algorithms used today, which can detect smaller matches too, – i.e. not only full documents and multi-page matches – are based on some kind of chunking method. During the analysis of the chunking methods I managed to correct one negative feature of two methods each by changing the algorithm.

Thesis 1

I created the partially overlapping word chunking which is the combination of word chunking and overlapping word chunking for plagiarism detection purposes, which I proved to be just as efficient in detecting similarities as the overlapping word chunking. At the same time depending on the implementation it requires a database of n-th size or the size of the query time becomes n times smaller (where n is the parameter of word chunking). [45]

Let us represent the number of words found in the document with *W*, the parameter of the chunking procedure with *n*, and the set of fragments with *Ch*. During the research I created a new chunking and query procedure, which I named as partially overlapping word chunking. The main point of this is that one of the documents (*q*) is processed by overlapping word chunking, while the other (*db*) is processed by word chunking and they are compared to each other. This solution eliminates the phase problem experienced at word chunking, because we create all possible pieces from one of the documents. We carry out the chunking of the other document on the other hand by using only word chunking, thus we have the possibility to reduce either the size of the database (~ $|Ch_{db}|$) to its n-th size: $|Ch_{db}| = W/n$, $|Ch_q| \approx W$, or else the searching time (~ $|Ch_q|$), that is the number of queries: $Ch_{db}| \approx W$, $|Chq| = W/n$

Using this chunking procedure insertion, deletion and substitution of one word can cause at most one mistake each, i.e only one fragment will be modified and the rest will still be evaluated the same by the system. This is exactly the same as the one experienced with overlapping word chunking, where there are *n* times more fragments but these errors effect *n* number of fragments, i.e. in the case of both procedures there must be one difference per each *n* word minimum for the system not to find a match.

Thesis 2

I proved about the overlapping hashed breakpoint chunking that by its use the text dependency of the hashed breakpoint chunking can be eliminated and thus it is suitable for the chunking of unknown texts in practice as well. [19][44][45]

I proved that the efficiency of the hashed breakpoint chunking procedure depends on the chosen hash code and the style and subject of the text. For this reason this algorithm cannot be efficiently used to detect similarities of texts of unknown origin or databases containing not homogeneous texts or ones of not the same subject or style. I pointed out that this text dependency can be eliminated by the parallel use of several hash values. This provides an opportunity to make a compromise depending on the application area between a higher and more uniform recall value and a larger database size.

Thesis 3

I created a new version of the n-gram algorithm – used in a wide circle nowadays to identify languages of documents – which cleans the final result of the earlier algorithm from the false positive findings deriving from the similarity of languages. The new algorithm is able to identify languages present in the text in a proportion larger than 30% even if these parts can be found in the text scattered and not in a coherent form. The new algorithm is suitable for filtering documents found in web corpuses and containing other languages in a quantity which already has a negative influence on plagiarism detection. [24]

For language identification one of the most frequently used algorithms is the n-gram algorithm, which has to scan the document only once in order to determine from the n-gram statistics which language the document was written in and – if proper samples are at our disposal – it can even determine its encoding. This algorithm can determine the most probable language of the text, however, in case of multilingual texts the second most probable language does not get the second lowest score. The reason for this is the similarity of the languages. I used this n-gram level similarity – not always matching the linguistic affinity – to filter the false positive findings from the similarity list.

The n-gram similarity in percentage (*h*) received by a document *D* should be h_1, h_2, h_3 in the descending order of the degree of the percent similarity, the languages should be indicated by L_1, L_2, L_3, i.e h_1 shows the similarity of document *D* to the patterns of the L_1 language, in percentage. The percent similarity between the languages will be indicated by h^{LiLk}. h_i' is the value given by the new algorithm to L_i language.

$$\mathrm{h_i' = h_i \quad if\ i=1}$$

$$\mathrm{h_i' = h_i - \frac{\sum_{k=1}^{i-1} h_k \times h^{LiLk}}{\sum_{k=1}^{i-1} h_i} \quad if\ i>1}$$

The algorithm, as a matter of fact, reduces the probability of all languages by the probability of the languages found before it, thus compensating for the distortion originating from the similarity between the languages.

Thesis 4

I created a new algorithm capable of finding cross-lingual plagiarisms, which breaks up the text into sentences instead of n-word chunking and compares the sentences with one another by means of a similarity metric imitating the process of translation to determine to what extent they are each other's translated versions. A significant feature of the new algorithm is that it does not require a machine translator but a dictionary is sufficient, which is easier to get and continuously develop. [1][11][23]

The basis of the algorithm is the bag of words model: a sentence (S_n) consisting n words are represented by the words inside ($w_1, w_2, ..., w_n$).

$$S_n^x = \{w_1^x, w_2^x, ..., w_n^x\}$$

The similarity (*Sim*) of two sentences is defined as follows:

$$\mathrm{Sim}(S_n^x, S_m^y) = \min\left(\alpha \cdot |S_n^x \cap S_m^y| - \beta \cdot |S_n^x \setminus S_m^y|, \alpha \cdot |S_m^y \cap S_n^x| - \beta \cdot |S_m^y \setminus S_n^x|\right)$$

Where α and β are constants used to weigh the common and missing words. Sim_a and Sim_b indicates the degree of similarity given by the similarity metric in case of the *a*-th and *b*-th sentences. Let us define two constants, two similarity degrees SIM_1 and SIM_2, where $SIM_1 < SIM_2$ and a distance value d. Based on the above we can write down that we consider the given document a match if the following is true:

- $\mathrm{Sim}_a \geq \mathrm{SIM}_2$ or
- $\mathrm{Sim}_a \geq \mathrm{SIM}_1$ and $\mathrm{Sim}_b \geq \mathrm{SIM}_1$ and $|a - b| < d$

The usability of the new similarity metric is stated in the next thesis.

Thesis 5

I showed that the new algorithm capable of finding cross-lingual plagiarisms can also be used in practice, has a result comparable with other algorithms, and that the value of the recall does not depend on the given pair of languages: in case of Hungarian-English its recall is 83%, while its precision is 40%, in case of German-English its recall is 83%, while its precision is 77% on the test corpus containing 12 Wikipedia-articles. [11][23]

To carry out the research I processed the whole English language Wikipedia containing about 4 million pages. I chose 12 articles from this at random, which were translated into Hungarian and German by translators. I showed that the new algorithm was able to find these and achieved a high recall – above 80% – on both texts and both languages. This is the sentence-level recall (*r*). The probability of the system finding at least one translated sentence from *k* sentences is: $1-(1-r)^k$, it can be easily seen that with the increase of the number of the sentences (*k*) this tends to 1, i.e.: it is more likely to find similar parts of a bigger size.

Thesis 6

I showed that the linear runtime of the new algorithm capable of finding cross-lingual plagiarisms in relation to the size of the database can be decreased to constant by using an information retrieval algorithm. [1][21][23]

I proved that the information retrieval algorithm that I use is suitable to filter out the sentences from a database of Wikipedia size which are most likely similar to the sentence searched for. The result of the experiment on the machine translated test corpus: in case of the Hungarian sentences the probability is 44% that the correct sentence is the first one, 13% that it is between the 2nd and 10th place and the probability of it being found between the 11th and 50th place is only 6% and it is 37% that it is not found between the first 50 (see figure 6.1)

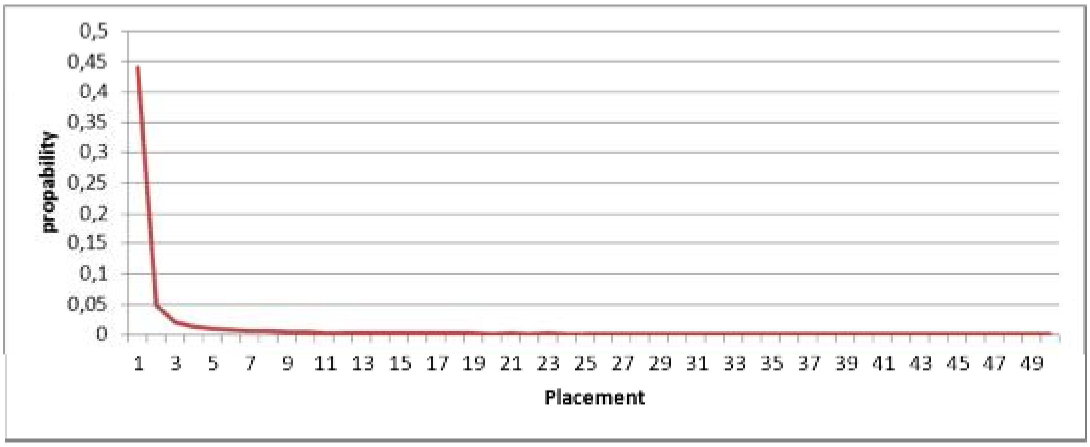

Figure 6.1.: Placement of correct results given back by indexed search (English-Hungarian)

I showed that the recall strongly depends on the length of the sentence due to the information retrieval algorithm: while the short sentences have a low recall value, the long and more meaningful sentences are more likely to be found (see figure 6.2).

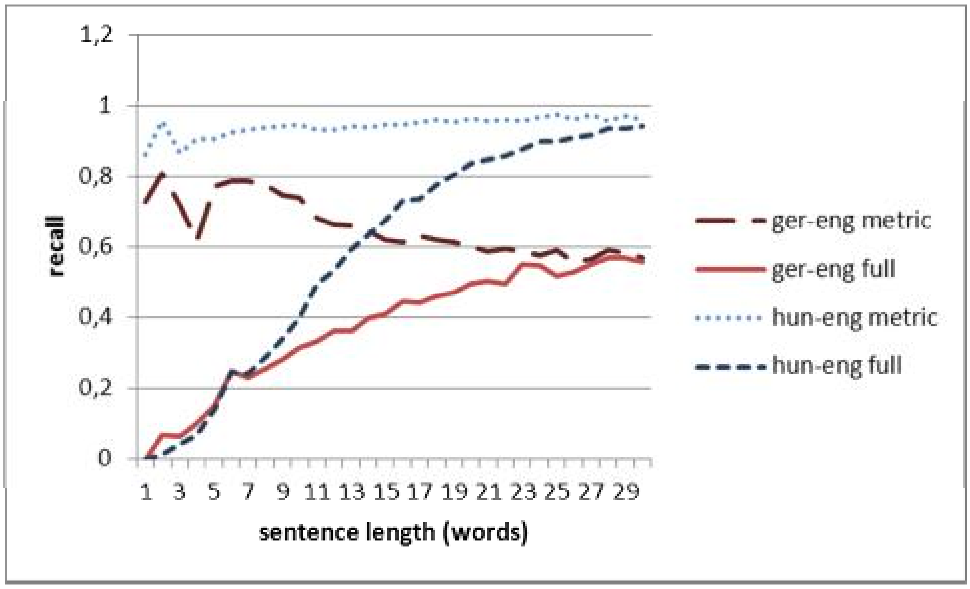

Figure 6.2: The value of recall for the similarity metric and the full system depending on the length of the sentence (for English-German and English- Hungarian language pairs)

By means of the information retrieval algorithm I was able to reduce the algorithms linear runtime in relation to the size of the database to constant, which makes the application of the algorithm in practice possible.

Thesis 7

I proved that my new algorithm capable of finding cross-lingual plagiarisms and based on a similarity metric is suitable for comparing not only translations but also texts written in the same language. In case of a text having been translated twice the sentence-level recall was 92% and 83%. This algorithm is completely insensitive to the word order within the sentence as compared to the n-gram based algorithm, where the change of the order of words results in the reduction of the recall. According to the sequencing of the results, the number of sentences and the value of the results on the test corpus containing the 12 Wikipedia-articles, the 11 correct results were put between the first 15 places, while the false positive findings were all but one present at the end of the results list.

The new algorithm used for detecting translational plagiarisms can even be used to compare monolingual texts. In this case – instead of a dictionary identity – synonym, antonym, hyponym and hypernym identities are introduced and we can evaluate the identity of two texts based on these: the intersection and difference of two bag of words in the similarity metric. All the other elements of the algorithm remained unchanged, I changed only the bilingual dictionary to the English WordNet. Plagiarizing can most simply be simulated by changing the words to synonyms, however, this would have been a too easy task due to the architecture of the algorithm. Therefore, I simulated the plagiarizing by having two machine translators translate the corpus containing the 12 Wikipedia-articles back from Hungarian to English, which were previously translated from English to Hungarian manually by translators. I tested the algorithm on these two corpuses containing serious alterations and mistakes. When using the first machine translator for the translated text it found 11 out of the 12 articles and the correct results were present among the first 15 places. I achieved a very similar result when using the second translator too, the 10 correct results were among the first 17 places.

Application of the results in practice

The overlapping word chunking has also been applied in practice in the SZTAKI KOPI Plagiarism Search Portal, ever since its start in 2004 this algorithm has ensured the base of the monolingual detection. The documents are stored in the database by word chunking and the system processes the documents to be compared by using overlapping word chunking, so by using the partially overlapping word chunking we have been able to make a smaller database.

The goal of my research on detecting cross-lingual plagiarisms was to find out whether it was possible – and if how efficiently – to identify translational plagiarisms between the English and the Hungarian language. Since the results are very promising and my new algorithm based on the similarity metric and the information retrieval algorithm has proved to be useful in practice too, the building in of this algorithm into the KOPI Plagiarism Search Engine has also become possible. At the end of year 2011 the KOPI Portal was the first in the world to provide a cross-lingual plagiarism detection service.

The primary goal of the KOPI Portal is to reduce plagiarizing in higher education. This is achieved by giving teachers an effective tool which makes plagiarizing risky. Being aware of the algorithms inside the KOPI system it must be admitted that hiding the traces of a plagiarism consumes much more energy than writing the given homework or dissertation honestly.

Acknowledgments

Isaac Newton said „If I have seen further, it is by standing on the shoulders of giants.", who when saying this was standing on the shoulders of Bernard of Chartres.

I would like to express my deepest gratitude to my parents for their support, and that they set a good example to follow. I am also very thankful to my wife and my father for the constant push and encouragement.

I would like to thank Krisztián Monostori that he introduced me more than 10 years ago to the field of plagiarism detection, Gábor Hodász for the first steps together in this field, Arkady Zaslavsky for the Australian scholarship where I could deepen my knowledge.

I am grateful to László Kovács that he invited me to work at MTA SZTAKI; András Micsik for his professional support; Zoltán Tóth for building KOPI with me; Miklós Vajna for the implementing and testing of the language identification algorithm; Balázs Pataki for the help during the implementation of the cross-lingual algorithm; Péter Pallinger for administering the software environments and systems for my research; Magdolna Zsivnovszki and Éva Virág for helping with my English (and Hungarian); Péter Inzelt for the MTA SZTAKI internal project support which made it possible to finish my research.

I am deeply grateful to my supervisor Gábor Prószéky for his support during my two years at Pázmány Péter Catholic University and the detailed, very thorough reviews.

List of Authors Publications

Journal Papers

[1] **Máté Pataki** and Attila Csaba Marosi, “Searching for Translated Plagiarism with the Help of Desktop Grids”, *Journal of Grid Computing*, Volume 11, Issue 1, pp 149-166, Springer, ISSN: 1570-7873, 2013.

[2] **Pataki Máté**, „Digitális könyvtárak védelme a KOPI plágiumkereső rendszerrel”, *Tudományos és Műszaki Tájékoztatás* 54/3., 2007.

[3] Kovács László és **Pataki Máté**, „E-ügyintézés bevezetése Kaposvárott”, *Jegyző és közigazgatás 8/1.*, ISSN: 1589-3383, 2006.

[4] **Pataki Máté**, „W3C ajánlások magyarul”, *Tudományos és Műszaki Tájékoztatás 52/9.*, 2005.

Books

[5] **Pataki Máté** és Abonyi-Tóth Andor, Szerkesztő: **Pataki Máté** „Bevezetés az info-kommunikációs akadálymentesítés világába II.”, ISBN: 978-615-5043-62-8, 2011.

[6] Abonyi-Tóth Andor, **Pataki Máté** és Mátételki Péter, Szerkesztő: Abonyi-Tóth Andor, „Bevezetés az info-kommunikációs akadálymentesítés világába I.”, ISBN: 978-615-5043-18-5, 2011.

[7] **Pataki Máté**, „Infokommunikációs akadályok”, ISBN: 978-615-5043-66-6, 2010.

Book Chapters

[8] Jókai Erika, Koloszár Kata, Mogánné Tölgyesy Szilvia és **Pataki Máté**, „Rehabilitációs támogató technológiák”, ISBN: 978-963-2790-97-8, 2010.

[9] Deákné Orosz Zsuzsa, Dr. Kecskeméti Éva, Zalabai Péterné, Abonyi-Tóth Andor, Fehérné Kovács Zsuzsa, Helfenbein Henrik és **Pataki Máté** „Fogyatékos személyek szociális segítése – Szociális ellátás”, ISBN: 987-615-5043-08-6, 2009.

[10] Dr. Kecskeméti Éva, dr. Nagy Janka Teodóra, Abonyi-Tóth Andor, Fehérné Kovács Zsuzsa, Földiné Angyalossy Zsuzsa, Helfenbein Henrik, dr. Márkus Eszter, **Pataki Máté**, dr. Perlusz Andrea és dr. Szabó Ákosné, „Esélyegyenlőség a joggyakorlatban - Szociális jogi szabályozás”, ISBN: 987-615-5043-10-9, 2009.

Foreign Conference Papers

[11] **Máté Pataki**, „A new approach for searching translated plagiarism", *5th International Plagiarism Conference*, Newcastle-Upon-Tyne, 2012.

[12] Hannes Eichner, András Micsik, **Máté Pataki** and Robert Woitsch, „A use case of service-based knowledge management for software development", *IFIP international conference on research and practical issues of enterprise information systems*, Győr, 2009.

[13] László Kovács and **Máté Pataki**, „Copy Protection via Plagiarism Search", *3rd International Plagiarism Conference*, Newcastle-Upon-Tyne, 2008.

[14] László Kovács, Zoltán Szentirmay, **Máté Pataki** and Péter Pallinger, „Development of the new National Cancer Registry", *microCAD 2007, International Scientific Conference*, ISBN: 978-963-661-759-2, Miskolc, 2007.

[15] László Kovács and **Máté Pataki**, „KOPI protection instead of copy protection", *Axmedis 2006,. 2nd International Conference on Automated Production of Cross Media Content for Multi-channel Distribution*, ISBN: 88-8453-526-3, Leeds, 2006.

[16] Roland Alton-Scheidl, András Micsik, **Máté Pataki**, Wolfgang Reutz, Jürgen Schmidt and Thomas Thurner, „StreamOnTheFly: a Peer-to-peer network for radio stations and podCasters", *Proceedings of the First International Conference on Automated Production of Cross Media Content for Multi-channel Distribution, AXMEDIS'05*, Florence, 2005.

[17] László Kovács, András Micsik, **Máté Pataki** and Robert Stachel, „StreamOnTheFly: a network for radio content dissemination", *Lecture Notes in Computer Science 3664*, 2005.

[18] **Máté Pataki**, „Plagiarism detection and document chunking methods", *Proceedings of the Twelfth International Conference on World Wide Web, WWW2003*, Budapest, 2003.

[19] Krisztián Monostori, Raphael Finkel, Arkady Zaslavsky, Gábor Hodász and **Máté Pataki**, „Comparison of Overlap Detection Techniques", *The 2002 International Conference on Computational Science, Lecture Notes in Computer Science 2329*, Amsterdam, 2002.

Hungarian Conference Papers

[20] **Pataki Máté**, Pataki Balázs, Tóth Zoltán, Pallinger Péter, Kovács László, „DRM megoldások áttekintése", *Networkshop*, Sopron, 2013.

[21] Micsik András, **Pataki Máté**, Garzó András, „A KOPI Plágiumkereső terhelésének elosztása cloud környezetben”, *Networkshop*, Sopron, 2013.

[22] **Pataki Máté**, „Algoritmus fordítások keresésére”, *BJMT Alkalmazott Matematikai Konferencia*, Győr, 2012.

[23] **Pataki Máté**, „Fordítási plágiumok keresése”, *MSZNY 2011. VIII. Magyar Számítógépes Nyelvészeti Konferencia*, ISBN: 978-963-306-121-3, Szeged, 2011.

[24] **Pataki Máté** és Vajna Miklós, „Többnyelvű dokumentum nyelvének megállapítása”, *MSZNY 2011. VIII. Magyar Számítógépes Nyelvészeti Konferencia*, ISBN: 978-963-306-121-3, Szeged, 2011.

[25] **Pataki Máté**, „Plágiumkeresés különböző nyelvek között”, *Networkshop*, Kaposvár, 2011.

[26] **Pataki Máté**, Abonyi-Tóth Andor és Helfenbein Henrik, „A "Bevezetés az esélyegyenlőséget szolgáló info-kommunikációs technológiákba" kurzus tapasztalatai az ELTE Informatikai Karán”, *III. Oktatás-Informatikai konferencia*, Tanulmánykötet, ISBN: 978-963-312-037-8, Budapest, 2011.

[27] **Pataki Máté**, „Plagizálás a felsőoktatásban”, *Magyar Tudomány Napja: Kövek, szavak, gondolatok – A kultúrák találkozása*, Szombathely, 2010.

[28] **Pataki Máté**, „Webes Akadálymentesítési Útmutató 2.0 - W3C WCAG 2.0”, *Akadálymentes web-tervezés workshop*, Veszprém, 2010.

[29] **Pataki Máté**, „Rámpát a honlapokra – úton az akadálymentes honlapok felé”, *Networkshop*, Szeged, 2009.

[30] **Pataki Máté** és Micsik András, „Üzleti modellen alapuló webes tudásprezentáció”, *Networkshop*, Szeged, 2009.

[31] **Pataki Máté**, Füzessy Tamás, Kovács László és Tóth, Zoltán, „Hibatűrő keresés digitalizált magyar nyelvű szövegekben”, *Networkshop*, Dunaújváros, 2008.

[32] **Pataki Máté**, Richter Viktor, „A W3C szabványosítási törekvései”, *Networkshop*, Dunaújváros, 2008.

[33] **Máté Pataki** and Tamás Füzessy, „Digitization errors in Hungarian documents”, *AACS '07 Automation and Applied Computer Science Workshop*, ISBN: 978-963-420-909-6, Budapest, 2007.

[34] **Pataki Máté**, Kovács László és Pataki Balázs, „Egy országos méretű orvosi adatbázissal kapcsolatos informatikai kihívások”, *Networkshop*, Eger, 2007.

[35] **Pataki Máté** és Tóth Zoltán, „Szkennelt szövegek digitalizálása során keletkező hibák elemzése magyar szövegek esetében", *Networkshop*, Eger, 2007.

[36] **Pataki Máté**, „A W3C és a Mobilweb", *Magyarországi Web Konferencia*, Budapest, 2007.

[37] **Pataki Máté**, „KOPI Plágiumkereső a digitális tartalmak védelmében", *DAT 2006 - A digitális kreatív iparágak szerepe Magyarországon*, Budapest, 2006.

[38] **Máté Pataki**, „Distributed similarity and plagiarism search", *AACS 2006, Proceedings of the Automation and Applied Computer Science Workshop*, ISBN: 963-420-865-7, Budapest, 2006.

[39] **Máté Pataki**, „Plagiarism search within one document", *AACS 2006, Proceedings of the Automation and Applied Computer Science Workshop*, ISBN: 963-420-865-7, Budapest, 2006.

[40] **Pataki Máté**, „W3C WAI, avagy weblapok akadálymentesítése", *Magyarországi Web Konferencia 2006*, Budapest, 2006.

[41] **Pataki Máté** és Kovács László, „W3C WAI - weblapok akadálymentesítése", *Networkshop*, Szeged, 2005.

[42] **Pataki Máté**, „KOPI online plágiumkereső és információs portál", *Networkshop*, Győr, 2004.

[43] Kézdi Tamás, Kovács László, Micsik András és **Pataki Máté**, „Elosztott digitális hangtárak a közösségi rádiózásért (SotF)", *Networkshop*, Pécs, 2003.

Other Publications

[44] **Pataki Máté**, „Szöveges dokumentumok darabolása és tömörítése hash-kódolással - darabolási technikák és másolatkeresés", *Budapesti Műszaki és Gazdaságtudományi Egyetem, diplomadolgozat*, Budapest, 2002.

[45] Hodász Gábor, **Pataki Máté**, „Szöveges dokumentumok darabolása és tömörítése", *Budapesti Műszaki és Gazdaságtudományi Egyetem, TDK dolgozat*, Budapest, 2001.