Wikipedia as Text

Máté Pataki, Miklós Vajna, Attila Csaba Marosi
Computer and Automation Research Institute - MTA SZTAKI
H-1518 Budapest, P.O. Box 63, Hungary
{mate.pataki, vajna, atisu}@sztaki.hu

When seeking information on the web Wikipedia is an essential source: its English version features nearly 4 million articles. Studies show that it is also the number one source of plagiarism, so when KOPI, a new translational plagiarism checker was created, a way to add this vast source of information to the database was to be found. As the whole database is impossible to be downloaded in an easy to handle format (like HTML or plain text) and all the available Mediawiki converters have some flaws, a Mediawiki XML dump to plain text converter has been written, which runs every time a new database dump appears on the site with the text version being published for everybody to use.

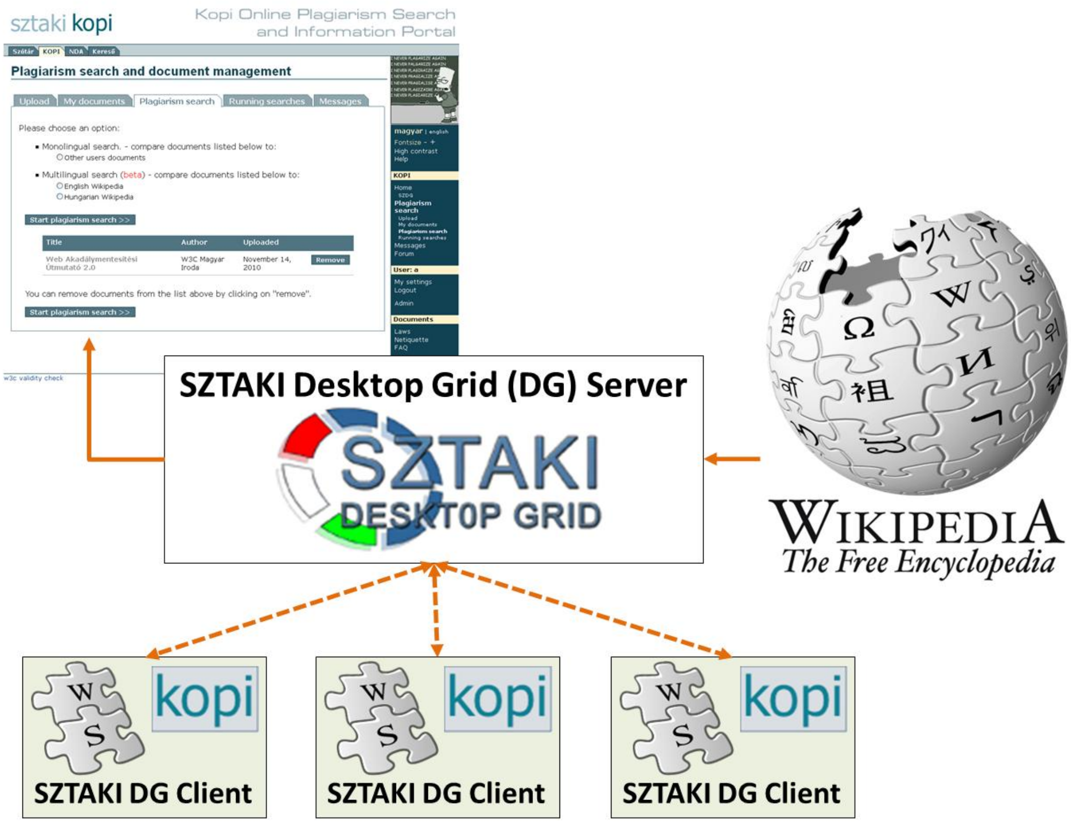
The KOPI Plagiarism Search Portal was developed by the Department of Distributed Systems (DSD) of MTA SZTAKI in 2004 as a plagiarism checker tool for carbon copy plagiarism cases. In 2010 the system improved by adding a unique feature to the search engine, the capability to find translated plagiarism. For this function it was necessary to include the whole English Wikipedia, as the number one source of potential plagiarism, into the database.

When the Wikipedia dumps were first downloaded from the server, all possibilities of converting them easily and quickly to plain text were examined, as plain text is easy to manipulate. Consequently, the whole content had to be run through a series of language processing steps. In most cases the available converters were not suitable for handling a larger chunk of XML dump, or the output was error prone and in many cases – when using, for example, a Mediawiki instance as a converter – the conversion was slow.

Considering these facts, a new converter needed to be written based on the following features:

- Article boundaries have to be kept
- Only the textual information is necessary
- Infoboxes – as they are duplicated information – are filtered out
- Comments, templates and math tags are dismissed
- Other pieces of information, like tables, are converted to text

Wikipedia dumps are published regularly and as the aim is to be up-to-date, the system needed an algorithm which is able to process the whole English Wikipedia in a fast and reliable way. As the text is also subject to a couple of language processing steps to facilitate plagiarism search, these steps were included and the whole processing moved to the SZTAKI Desktop Grid service (operated by the Laboratory of Parallel and Distributed Systems), where more than 40,000 users have donated their free computational resources to scientific and social issues. Desktop grids are usually suited for parameter study or “bag-of-tasks” type of applications and have other minor requirements for the applications in exchange for the large amount of “free” computing resources made available through them. SZTAKI Desktop Grid was established in 2005 and it utilizes mostly volatile and non-dedicated resources (usually the donated computation time of desktop computers) to solve compute intensive tasks from different scientific domains like mathematics and physics. Donors run a lightweight client in the background which downloads and executes tasks. The client also makes sure that only the excess resources are utilized so there is no slowdown for the computer and the donor is not affected in any other way.



sztaki kopi

Kopi Online Plagiarism Search and Information Portal

Plagiarism search and document management

Please choose an option:

- Monolingual search. - compare documents listed below to:
  ○ Other users documents
- Multilingual search (beta) - compare documents listed below to:
  ○ English Wikipedia
  ○ Hungarian Wikipedia

Title	Author	Uploaded	
Web Akadálymentesitési Útmutató 2.0	W3C Magyar Iroda	November 14, 2010	Remove

You can remove documents from the list above by clicking on "remove".

Wikipedia to text conversion process

The Mediawiki converter was written in PHP to support easy development and compatibility with the existing codebase of the KOPI Portal. The main functionality could be implemented with less than 400 lines of code. The result was adapted to the requirements of the desktop grid with the help of GenWrapper, a framework specially created for porting existing scientific applications to desktop grids. GenWrapper's primary goal is lowering the porting effort required for so called legacy applications, which either have no source code available or making changes to them is infeasible. In case of KOPI it allowed developing the converter independently from the desktop grid and the result could be effortlessly deployed.

This new arrangement allows Wikipedia in any language to be converted and preprocessed in a couple of days. As these text versions can be used for several other purposes as well, they are shared and made available to everybody. Currently one can download the English (5,7 GB), Hungarian (300 MB), German (2,2 GB) and French (1,5 GB) versions (sizes are gz compressed size). Based on the project work and plans, other languages will follow shortly.

Links:

Download link for Wikipedia as text
http://kopiwiki.dsd.sztaki.hu
KOPI portal
http://kopi.sztaki.hu
Department of Distributed Systems of MTA SZTAKI
http://dsd.sztaki.hu
SZTAKI Desktop Grid
http://szdg.lpds.sztaki.hu
GenWrapper
http://genwrapper.sourceforge.net
Laboratory of Parallel and Distributed System
http://www.lpds.sztaki.hu

Please contact:

Máté Pataki, MTA SZTAKI, Budapest, Hungary
Tel: +36 1 279 6269
E-mail: mate.pataki@sztaki.hu