DIADEM – the next step for web search

Searching the internet for complex information is set to become far easier, if a major new research project at Oxford University succeeds.

The Diadem project, which has just been selected to receive a €2.4m grant from the European Research Council, has its sights set on one of the knottiest problems, and most potentially significant advances, in web search technology.

Headed by Georg Gottlob, Professor of Computing Science at Oxford University, the five-year programme aims to create software that can automatically examine websites in a certain subject area, understand their structure, retrieve useful information and present it clearly to the user.

This will allow web users to search the 'deep web' – the specialist databases that contain much of the internet's information. Companies like Google, Yahoo! and Microsoft are keenly interested in this possibility, which would let them create the next generation of search engines.

The current generation relies on searching with keywords. This works well for some queries – typing 'pasta al pesto recipe' into a search engine quickly supplies the information needed. But searching for 'restaurants near me serving pasta al pesto as today's special' will turn up only vast quantities of irrelevant information.

To find what you're looking for, you'll need to spend a long time looking through individual websites, each with a different design and structure. Humans find this kind of task easy, if boring – we can visit a web page we've never seen before and immediately understand which information is where. But computers have trouble reading this kind of semi-structured content – they have no way of telling which section of a page is a menu, or which of the numbers onscreen is an item's price, for example.

All the important information is there, but the computer can't read it, because it doesn't understand how it is structured. Diadem – Domain-centric Intelligent Automated Data Extraction Methodology – aims to change this. Gottlob envisages software that will let computers pull highly structured information accurately out of the chaos of the Internet, learning each website's structure as they go.

If the project team succeeds, in five years it will have built such software. One goal is a system to analyse a specified country's property market. For example, the UK Yellow Pages lists some 17,000 estate agents and their websites; once supplied with their URLs, the software will go through all of them, extract information and display it clearly and conveniently or store it in a database.

But Diadem's results could equally apply to other domains – the researchers aim to create a toolkit to let others easily build versions to search the websites of restaurants, schools, travel agents, airlines, retailers or any other area they want to find out about.

To do this, Gottlob and his team will use an innovative 'knowledge-based' approach, whereby the software combines high-level knowledge about how a domain like the property market works with low-level facts it picks up as it analyses the web pages in it.

Unlike current search engines, which look for words on web pages, Diadem will analyse the code behind them, understanding how it translates into objects on the page – a growing trend known as 'object search'. Diadem brings this together with another hot topic – 'vertical search', or search engines that specialise in a particular subject area.

When set loose on a new domain, the software will firstly analyse the websites in it. The knowledge this gives it will be used to create highly efficient programmes to extract information from each one. Finally, these programmes will be run in parallel for many websites in the domain at the same time using 'cloud computing', in which tasks are sent to powerful computers distributed over the Internet. The result should be a rich database reproducing the information extracted from each website in structured form, which users can readily manipulate and analyse.

Others have made steps in similar directions, but previous technologies haven't aimed as high as Diadem. Some need to be taught to read each website they encounter by a human operator – Diadem aims to produce software that can be told broadly about what to look out for in a particular domain and then left on its own to learn how to find it in each website.

Gottlob himself has already had success in this area; he is a founder of Lixto Software, a Vienna-based company specialising in automated extraction of data from semi-structured webpages for purposes that range from monitoring competitors' prices to business process integration. Its clients include major travel agencies and automotive firms.

But Diadem's goals go much further. If it succeeds, the project will revolutionise our ability to extract useful information from the net, moving us closer to the holy grail of the 'semantic web' – an internet on which every webpage contains detailed information instructing computers on its structure and content. This could have huge economic value – leading search engine firms have already expressed interest in Diadem's results.

It's an extremely ambitious goal, and there will be major challenges along the way. Gottlob and his team will have to make several technological breakthroughs. Crucially, they'll need to devise powerful new ways of querying and analysing websites that can wade through thousands of pages containing countless forms, menus and data fields, and learn how they work on the fly, with no human input needed.

The rewards could be great, though. The dominant search engine companies are fully aware that current search technologies are not enough, and that a paradigm shift is needed to give access to the web's hidden wealth of information. Even apart from the benefits to web users everywhere, such a breakthrough would almost certainly prove very lucrative.