Object Muscat, an Open Source search engine

Martin Porter, Richard Boulton
Bright Station PLC.

A few people in the IR community will already have heard of the Muscat software. It was begun in the early 1980s by Martin Porter, as a generalised and self-contained piece of software for managing computerised catalogues, and for creating Information Retrieval systems. It was written with library and museum catalogues very much in mind, although it was found to be readily applicable to other kinds of data. Porter had been much influenced by contact with Keith van Rijsbergen and Stephen Robertson, with whom he worked in Cambridge on a two-year research project in 1979-81, and subsequently tried to capture in Muscat a faithful implementation of the probability theory model of IR, which then and subsequently has been so closely linked with Stephen Robertson's name.

As a commercial system, Muscat has enjoyed a reasonable success. It is fast, flexible and scalable. It can structure and index documents in a bewildering variety of ways. It will comfortably handle an IR system with a couple of thousand documents, as well as one with over 100 million. To the IR community it has been of interest, but, as a tool for testing out and demonstrating IR ideas, not especially useful, certainly not useful in the way that Okapi has been.

The Muscat software used to belong to Muscat Ltd, but Muscat Ltd has now been entirely absorbed in Dialog (now renamed as Bright Station). A need for a rewrite of the Muscat software has been apparent for long time.
It is unnecessary to explain why: suffice it to say that the code goes back to the early 1980s, and it is quite remarkable that it has lasted so long. Last year Dialog finally made the decision to back a major new round of software development, and in doing so made a bold choice. It was decided that the new software should be open sourced, under the GPL license agreement.

Just to add to the confusion, it was decided to call the new software "Muscat" as well! Throughout the rest of this note, the new software will be referred to as Object Muscat, in contrast to old Muscat, the system developed over a 20 year period, up to the end of last year.

Object Muscat is the work of a small team of software developers based in Cambridge, and being led by Richard Boulton. The official web address for the work is,

http://open.muscat.com/

We have discovered that we are not the only people open sourcing search engines. The interested reader may be referred to projects such as

- ht://Dig (http://www.htdig.org/) from San Diego State University
- Swish++ (http://www.best.com/~pjl/software/swish/), and
- Senga (http://www.senga.org/)

Nevertheless, we believe Object Muscat has a good chance of becoming the de facto open source search engine as part of the wealth of software which comprises the open source revolution. There are a number of reasons for this:

We can build upon experience gained, and lessons learnt, from old Muscat. To give a simple example, old Muscat did not keep a mapping D->L(D) of document identifier D to its normalised document length L(D). It must be remembered that normalised document length did not play a part in the probabilistic formulae of the early 1980s. Its use was never demanded by customers, who would not have understood its significance, and with the passage of time and the elaboration of the various secondary storage structures it became increasingly difficult to add in. The need for it is now apparent, and it will of course be a standard part of Object Muscat. We can also benchmark Object Muscat's IR speeds against old Muscat, which is recognised as being good. For simple queries Object Muscat is equally good, for large probabilistic queries it can outperform old Muscat by a factor of 3 or 4.

We can test Object Muscat inside our parent company which is driven by a need for good IR software, as well as outside it. We are already replacing old Muscat with Object Muscat inside WebTop, a global search engine which Dialog is developing (http://www.webtop.com/). This is an excellent test of its robustness and performance, and of new software to support distributed searching.

We are still very keen to follow the probabilistic model, which of course is known to perform well. By default, Object Muscat uses the BM25 weighting scheme. Later this year we expect to get some hard evaluations with the retrospective TREC data. So we hope to create a search engine that is as close to being state-of-the-art in IR performance as we can make it.

Our hope with Object Muscat is that it will be useful to the IR research community in a way that old Muscat clearly never was. The open sourcing of the code is of course a huge benefit here. Anything we create can be distributed to academia free of charge under the GPL license. Anything academia wishes to add can be similarly distributed. The old business models for developing software inevitably led to mutual suspicion between the business and academic worlds. To the academics the business people stole their ideas and squabbled among themselves over IPR. To the business people the academics were after funds for pure research, while failing to appreciate the harsh realities of business life. In the field of IR, where no significant advance has ever been made outside academia, this has been very damaging. The open source model for developing IR systems blows all that away.

So far indeed the signs are encouraging. In March 2000 we had over 700 people accessing our Web site, and 100 separate downloads of the open source code. We have had a steady stream of Email and phone enquiries, and are probably going to work with a couple of Departments in UK Universities on IR projects.