



### FAST, ROBUST, SCALABLE SEARCH – WITH PROBABILITY

Third party evaluation of Grapeshot has indicated it is very fast – processing queries across millions of documents in 4 milliseconds, on standard home computing hardware. Grapeshot is a technology designed for the implicit search infrastructure required to power new business intelligence and contextual advertising applications. Grapeshot is licensed as an OEM technology for software application designers to build within their own software stack.

### TECHNOLOGY USP's

Grapeshot first revealed its WordRank methods to Google in 2005, following an approach by Google to Dr. Martin Porter, Grapeshot's founder, for permission to use some of his algorithms. Google's Head of Search Quality commented that Grapeshot had three unique features:

1. **Multiple-word querying:** whole documents (400 plus words) or excerpts from emails or browser pages can be used as the search input, with multiple word queries processed in milliseconds. (Google had a limit of 12 words).
2. **Dynamic word suggestion:** like a folksonomy or taxonomy, Grapeshot suggests useful related terms with no upfront taxonomy required, and no overhead of using people to create semantic rules. This means Grapeshot effortlessly scales to many languages including Chinese, Japanese and Arabic.
3. **Variable word-weights:** the core of WordRank which offers personalization and auto-categorization opportunities. Word-weights change per user – advancing the ability to segment an audience.

### THE GRAPESHOT VALUE

The use of probabilistic models in the Grapeshot algorithms is vital to provide an adaptive learning capacity. Whereas Google uses "PageRank" to rank documents, Grapeshot uses new algorithms to achieve "WordRank". Essentially every word (or token) is given a weight of significance. This is not based on word frequency inside documents – as that only delivers a search engine that always gives the same answer to the same word typed in the search box.

No, Grapeshot has a rather more elaborate method of seeing the distribution of each word across the whole corpus, and making Bayesian inference about the word weight for that one specific user. Word weights help to establish the best N terms per document, and are used to heighten high precision search, as against just high recall. More significantly, when a user looks at one document – the word weights of all words in that document change, only for that one user. It means there is a spectrum of weights attached to any one word, right across the population of users. Suddenly the same word "ipod" featuring inside the content or meta-data category on an XML document, can have different weights for different users, thereby starting to provide a personal definition of relevance to any one user, relative to the rest of the audience.

Grapeshot can log the word weights for each user, which act as a "concept cloud" around the user, that then helps to skew the delivery of new documents from a publisher's content management system. Therefore if I am interested in "iTunes", "ipods" and have the words already weighted relatively high in my personal profile,



then any new data that includes "ipods" (as a text word, or audio/video sequence token) will be advanced to me ahead of other users who do not feature such high weights (probability-wise) within their particular profiles. Note that profiles can be hidden (implicit) or available to be seen by the user (explicit).

Grapeshot's unique "WordRank" algorithms can attribute variable word weights to each and every user. With the advantage of a spectrum of word weights, per user, for any given word; Grapeshot can now advance advertisements or content to the top 10% or 25% percentiles of the spectrum range - offering enhanced targeting of content (advertising and business intelligence) or a personalized content experience (for the publisher's "know your customer and audience" agenda).

### BACKGROUND

Grapeshot's founder, Dr. Martin Porter, is the Cambridge University computer scientist who developed the language stemming techniques known as the Porter Stemmer. His stemmers take the suffix endings off words, such that "connections", "connecting" and "connects" all become "connect" for the purposes of retrieval. Likewise "avoir", "être" and other irregular verbs need to be accommodated. Dr. Porter has published language stemmers for 12 European languages, including English, French, Spanish, Portuguese, Italian, German, Dutch, Swedish, Norwegian, Danish, Russian, and Finnish. Dr. Porter makes his stemmers available as open source through his [Snowball](#) website. Microsoft, Google, IBM and other corporations have made personal contact with Dr. Porter to request the use of his work in their products and services to date. His academic paper (1980) An algorithm for suffix stripping. *Program*, 14 :130–137 now has over 800 academic cross-citations, and is one of the most popular papers in the field of Information Retrieval. Dr. Porter's made an "outstanding contribution to the field of information retrieval" and won the [Strix Award](#) in recognition for his work.

In 1992 Dr. Porter started to commercialize his University work, selling the MUSEum CATaloguing system called Muscat as a new internet search technology. Muscat powered the first European internet search engine called Euroferret (1994) and provided search engines for the main BBC website, Reuters, Nokia, Shell and other large corporations. Due to the success of the Muscat business, the technology was sold in 1997 for \$15 million to the Dialog Corporation (now part of the Thomson financial publishing group) which sold business news and intelligence to corporations worldwide.

In 1999 Dr Porter contributed new work to the creation of WebTop, a 500million document index of the internet - at the same time that Google had a similar sized internet search service. Alta-Vista only indexed 300m documents at the time. This technical achievement meant innovating with distributed search and learning how to index millions of documents using a few low specification Linux machines, yet still deliver sub-second response times. Due to internal politics within the Dialog Corporation, WebTop's parent company, the global index was not continued as a project, despite its technical innovations.

In 2001 both John Snyder and Dr. Porter resigned, and commenced work on Grapeshot - the fifth search engine that Dr. Porter has now written. Grapeshot has an XML data design and achieves a fast Bayesian probabilistic information retrieval search, using just 300k of code footprint. This means the software is highly portable, yet able to search millions of XML records at sub-second speeds - an ideal candidate for mobile and set-top box implementation. The Grapeshot algorithms introduce opportunities for personalization of a user's content experience, and allow publishers to capture the conceptual essence of their audience.



#### CALL TO ACTION

Grapeshot can be seen in action at <http://search.doodle.com>

Passwords for Doodle are available on request from [john.snyder@grapeshot.co.uk](mailto:john.snyder@grapeshot.co.uk)

Grapeshot's website - written up for the conversations with Google - is at: [www.grapeshot.co.uk](http://www.grapeshot.co.uk)