gSeek – An Autonomous Information Cataloguing Service

Mohit Jain[1], **Siddhartha Lal**[1], **P. Sai Teja**[1], **Sai Gopal Thota**[1], **Satyajit Swain**[1]

[1]Dhirubhai Ambani Institute of Information and Communication Technology
Post Bag No. 4 Near Indroda Circle
Gandhinagar 382 007 Gujarat (India)

{miohit_jain, siddhartha_lal, sai_teja, thota_gopal, satyajit_swain}@daiict.ac.in

Abstract

"It makes all the difference whether one sees darkness through light or brightness through the shadows."

– David Lindsay

Believing in the aforementioned adage, gSeek seeks to bring light to darkness from the vast sea of information that we all know as the internet.

We have often felt the need for information to be available to us before we ask for it. For information to be available to us as, when, and how we want it. Most importantly, we have felt the need for various *avatars* of information to be available to us at one place, catalogued properly. Searching the web for content that is similar *semantically* can be quite a cumbersome job, with the poor user required to go to umpteen number of search pages before he gets the information he seeks. We make the case that gSeek can target these problems and offer a new and enriching experience to the user.

Keywords

Information Cataloguing, Semantics, Web 2.0, Handwriting Recognition, Keyword Extraction, Query Processing

Motivation

We have often been irritated by the number of search pages we have to visit in order to gather semantically linked information about a subject. We would like to have all kinds of search results displayed at one place; be it text, audio, images, video, podcasts, code or books. We shouldn't need to look for them in different places. gSeek offers a solution to this problem in a jiffy.

Imagine another scenario. A research scholar is writing a paper on his research topic. He would want information, pertaining to this research area, to come to him rather than him going to it. The nature of this information can be quite varied. The scholar might be looking for related research publications. He might be looking for images, or even lectures (in audio or video format) on the same subject by a professor teaching in some faraway university. More importantly, the semantics of this information could change with time. By continuously scanning the researcher's notes and providing only relevant information, gSeek offers a solution to this predicament.

Another case in point can be a group of students working on a course project. They would want their work to be available to everyone in the group. Most importantly, Larry would want search results that *he* deems important to be shared with the rest of his group mates. Web 2.0 offers the opportunity for collaborative learning and gSeek proposes to make use of this prospect.

What gSeek can do?

Larry is a research scholar at the University of Timbuktu. His research area is Computer Networks. He takes a photo or scanned copy of his notes and makes them available to gSeek. If the notes were made on a Desktop PC or a PDA, then the task is even simpler.

gSeek then reads through the notes[1], identifies semantically relevant keywords[2] and performs contextual key searches for these on the internet. It tags all the relevant search result content; be it text, images, audio or video, along with these notes. The user can choose to make use of it as and when he likes it. gSeek also links up the current information and notes

---

[1] An image file undergoes hand writing recognition. Handwriting recognition can be achieved using the *Algorithm Segmentation based approach for lexicon environment* [by Bozinovic & Srihari (1989)] or by using *Offline Cursive Handwriting Recognition System based on Hybrid Markov Model and Neural Networks* [by Yong Haw Tay, Marzuki Khalid, Rubiyah Yusof, and C. Viard-Gaudin]

[2] gSeek identifies keywords using WordNet/TFIFD scores. Keywords can also be generated using techniques described in *Layout and Content Extraction for PDF Documents* [By Hui Chao and Jian Fan] and *Acquisition of Categorized Named Entities for Web Search* [by Marius Pasca]

with the previous nuggets of information that the user has made use of and intelligently correlates[3] them.

gSeek even enables online collaboration between Larry and his research group. By making use of the web services incorporated in the system, data can flow between the user and his network of people via some central server.

gSeek provides Larry with the option of archiving links that *he* deems important along with the respective documents. Along with these links, he is shown newer links as the document changes over time. When Larry chooses to share his work with his peers, these links prop up alongside and can be browsed by his group mates.

The content also becomes more focused as on the user uses the software overtime and filters out unnecessary blocks of information and gives him what he desires. The software reads through the notes and creates a dictionary of keywords for linking the users' notes. This dictionary of keywords also serves to create contexts and content for the automated search. The user can restructure his notes anytime and when he does so, gSeek updates its database accordingly.

Use Cases

1. Larry plans to write a research paper on the disadvantages of IEEE 802.5 protocol. The material available in books and on the net is humungous. He takes photographs of his notes and textbook and uploads them into gSeek. He can even make notes on his Desktop PC or PDA and make them available to gSeek.

2. gSeek now 'reads through' the notes, identifies keywords and conducts contextual key search on the net. It then tags all the relevant search result content - text, images audio, video etc and makes them available to Larry at one place. Apart from these, he even gets RSS feeds and podcasts on recent happenings in his research field.

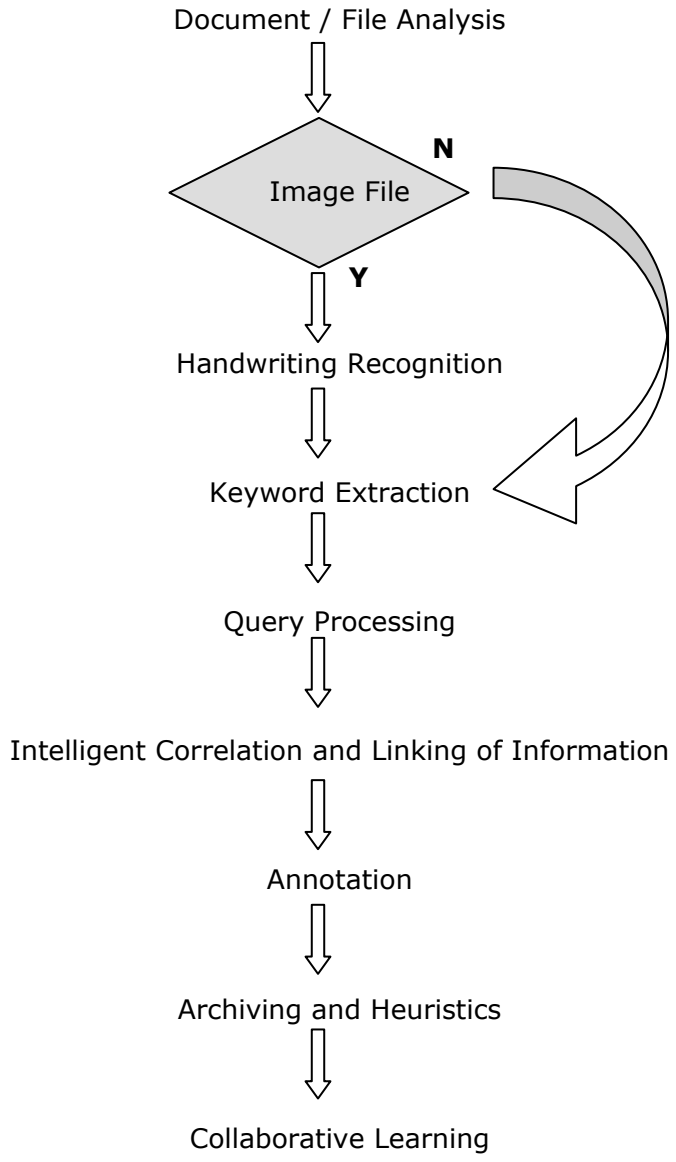
3. Larry then skims through the links and documents, rates them and discards a few irrelevant ones while also appending his own annotations wherever he wants. He is amazed at what the software can do. He then uses the software for further research on his topic and makes/uploads new notes into the software.

4. The software now links up the current notes with previous chunks of information. Over time, the content becomes more focused. Larry can also choose to share his notes with his peer group and get inputs from them. He can also go through their notes and bookmark links that they deem to be important.

5. Larry can not only use gSeek for work, but also to make his browsing experience a lot simpler. He types in a query for *Turing Test* in the search box and hits enter. He is immediately shown the top Google results for his query. But apart from this, he gets links to research publications pertaining to the query along with podcasts of a lecture that Dr Who delivered some days ago.

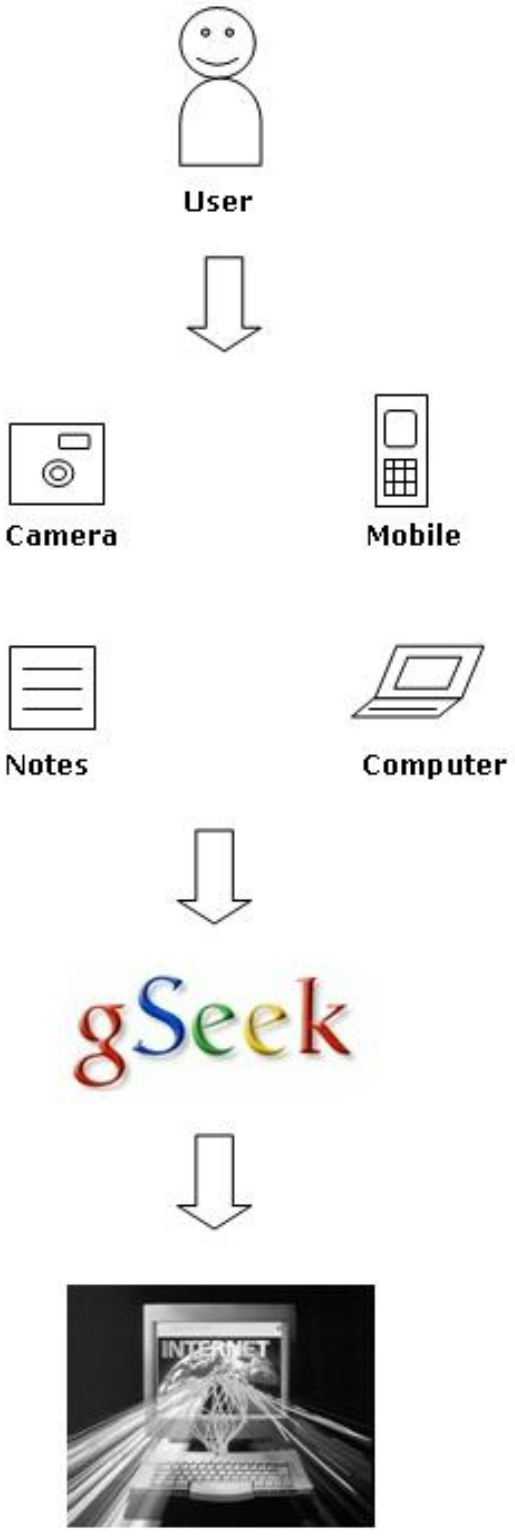
This online collaborative learning not only enhances his understanding of the topic, but also makes the learning experience a lot less cumbersome.

Application Workflow



[3] Collected information can undergo intelligent correlation and linking using techniques described in *Automated Discovery of Dependencies Between Logical Components in Document Image Understanding* [by Donato Malerba, Floriana Esposi, Francesca A. Lisi and Oronzo Altamura] and *Integrating Probabilistic Extraction Models and Data Mining to Discover Relations and Patterns in Text* [by Aron Culotta, Andrew McCallum and Jonathan Betz]

System Design



Security Implications

No special security measures need be taken other than those provided to services like Gmail, Orkut, Google Docs and personalized Google webpage - iGoogle. Moreover, if the nature of the information is confidential, then the user can choose not to share the document with his peers. Similarly, browsed links can have restricted access, with peers not being allowed to go through them.

Limitations

1. Number of file formats being accepted is limited i.e text and handwritten notes.
2. There is a space constraint if the files are uploaded onto the web as there might be limited webspace for a particular user profile.
3. There is no foolproof algorithm for extraction of keywords from text documents. Hence the user may be shown irrelevant results, not conforming to his requirements.
4. Handwriting-to-text conversion of scanned notes might introduce some errors into the queries being processed by the application.
5. Content Based Retrieval of Audio, Video or Image data is not supported.

Privacy Issues

1. The application would be on the user's desktop and hence all the notes and documents would be private.
2. The document will not be shared with the user's peers unless he explicitly asks for it.
3. The people with whom the document can be shared can be specified by the user. E.g. members in his study group, all or none
4. The documents won't be scanned by Google unless the user chooses to upload it and share it with other gSeek users.
5. References (search results) used by the user can be blocked or shown to others based on user's approval.
6. If the person likes to protect his information from others but would still like to use the facilities provided by gSeek, he can do so by utilizing the desktop application as an interface to the web.

Monetization

Every PC user is a prospective customer. The audience for gSeek is anyone who uses a computer or a mobile and likes to preserve and organize information.

Students can use it as an active learning aid; researchers and academicians can use it to organize relevant information; while business men and professionals can use it to keep track of latest happenings in their field of interest.

The user can also get online ads by allowing Google AdSense to go through the browsed and archived links. These advertisements would be more in tune with the kind of services and products that the user seeks; thereby increasing his possibility of going through them and making use of them.

The product is one of its kind as it fuses the traditional modes of information searching and information cataloguing with the myriad array of services provided by the web.

Potential for Extensibility

1. The algorithm chosen for keyword extraction can be improved further.

2. Audio and Video content can be made available to gSeek and relevant information can be searched for on the web by making use of Content Based Retrieval.

3. The product can be optimized so that it can be used on mobile phones and PDAs as well.

4. AI and heuristic techniques can be used to refine search with every single click.

Justification

Can scraps of paper that I write, lectures that I listen to, and pictures that I store; combined with select information from the deluge of that available on the web, serve as powerful and meaningful data to me? Can this information come to me before my asking for it? Can this information be linked *semantically* instead of being linked *syntactically*?

While managing files we all have queries and loads of doubts. Like when we are reading a document we come across several terms that we want to know about. Similarly, while listening to a song, we would like to know about the band and their compositions; fans would like to download images of the band members.

But as we are living in a fast-moving world, we don't find the time and patience to go to different webpages and look for relevant information. What if all this information could come to us before we even ask for it? What if all this could be linked and catalogued in a manner we want it to be? What if all this could be stored so that we could go through it whenever we have the time to do so?

gSeek does exactly this and fills the gap between you and your Google search; cataloguing all sorts of information for your perusal. It does before you ask it to do so. Not only this, gSeek tags relevant information (links visited by a user) alongside class notes or a document so that our scholar can go through them at a later point of time. This allows a group of peers to work together, building upon a common bank of information.

Inspired by the vision of Semantic Web, Web 2.0, and Collaborative Learning, gSeek proposes to answer these questions and seeks to make information searching an enriching experience. It proposes to make learning less tedious so that information can flow from one entity to another in a seamless fashion. More importantly, it seeks to convert information into knowledge and make the transition less cumbersome for the quintessential knowledge seeker.

References

[1] Marius Pasca

Acquisition of Categorized Named Entities for Web Search

[2] Donato Malerba, Floriana Esposi, Francesca A. Lisi and Oronzo Altamura

Automated Discovery of Dependencies Between Logical Components in Document Image Understanding

[3] Yevgen Biletskiy, Olga Vorochek, and Alexander Medovoy

Building Ontologies for Interoperability among Learning Objects and Learners

[4] Aron Culotta, Andrew McCallum and Jonathan Betz

Integrating Probabilistic Extraction Models and Data Mining to Discover Relations and Patterns in Text

[5] Giovanni Semeraro, Stefano Ferilli, Nicola Fanizzi, and Floriana Esposito

Document Classification and Interpretation through the Inference of Logic-Based Models

[6] Chichang Jou and Hung-Chang Lee

Handwritten Numeral Recognition Based on Simplified Feature Extraction, Structural Classification, and Fuzzy Memberships

[7] Hui Chao and Jian Fan

Layout and Content Extraction for PDF Documents

[8] Yong Haw Tay, Marzuki Khalid, Rubiyah Yusof, and C. Viard-Gaudin

Offline Cursive Handwriting Recognition System based on Hybrid Markov Model and Neural Networks