COMPANY RETRIEVAL SYSTEM

Assignment 2: Project Report

MULTIMEDIA RETRIEVAL COMP5425

Submitted by:
Lingzuo Zhao (lzha9833@uni.sydney.edu.au)
Nikhilesh Sharma (nsha4326@uni.sydney.edu.au)

Contents

Background

With the continuous development of technologies such as the internet and the web, various applications based on the internet and web technologies are emerging in an endless stream for meeting the needs of people in life and work. For example, some websites can provide the latest news about high-tech electronic devices like smartphones, cameras, audio and video devices etc.

The public are interested in these high-tech products, and try to buy the most cost-effective items to keep up with the current mass lifestyle and extend usage time as much as possible or to buy the most cutting-edge ones to experience exceptional functions before others.

These hopes are highly related to the company that offers the products, the company's financial situation, investment plan, relationship with parts suppliers and distributors, intellectual property about the technology involved in the products and services. All of these contribute to the products' survival time and the products' sustainable development in the market.

But the problem arises due to the fact that the data is distributed and the user has to search through multiple websites to gain information about a company and/or their products. There is a need for system that can centralize all this information and present it to the user. Some of the different types information that users are interested in, include financial information of companies and latest news prevailing in the market about those companies. Users are also interested in gathering information about various products and their specifications, companies have on offer.

Motivation

Undertaking this project and implementing the idea of having a centralized product information system was motivated by the fact that a user must spend a lot of time browsing the internet in order to obtain the desire information. Our project aims to provide the user with a centralized system that displays the relevant information about a company and their products. This removes the laborious effort for browsing the internet to find all the desired information.

Related Works

There are some websites on the internet that provide an interface for users to search and retrieve information about different companies [5]. CrunchBase is a database of the start-up ecosystem consisting of investors, incubators and start-ups, which comprises around 500,000 data points profiling companies, people, funds and events. The company claims to have more than 50,000 active contributors. Members of the public, subject to registration, can make submissions to the database; however, all changes are subject to review by a moderator before being accepted. Data is constantly reviewed by editors to ensure it is up to date. In 2013, CrunchBase said it had 2 million users accessing its database each month. [1]

Owler is an American internet company that crowdsources business insights by providing news alerts, company profiles, and polls and allows members to follow, track, and research companies in real time [2].

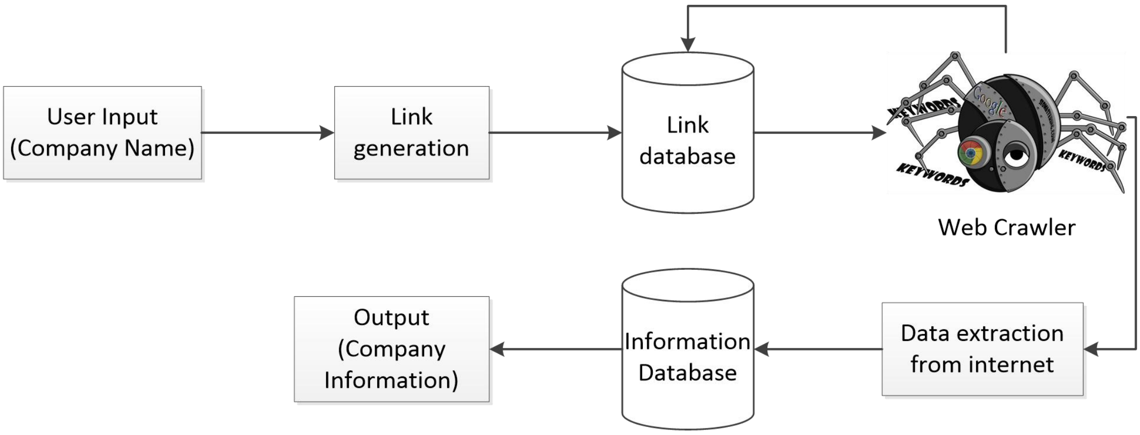
Many other similar websites exist that provide the following features to the user:

- A full database of private and public companies
- Curated, company-specific news
- Revenue and employee estimates
- Funding, acquisitions, and leadership change alerts

- Exclusive crowd sentiment polls measuring a company's perceived value

System Design

The complete pipeline of the system is given the flowchart below:



The designed system can be broken down into the following parts and steps:

- **User Input:**
  The user input is a windows form in which the user can input the name of the company he/she wishes to obtain information about. The figure below is an example of the user input form.

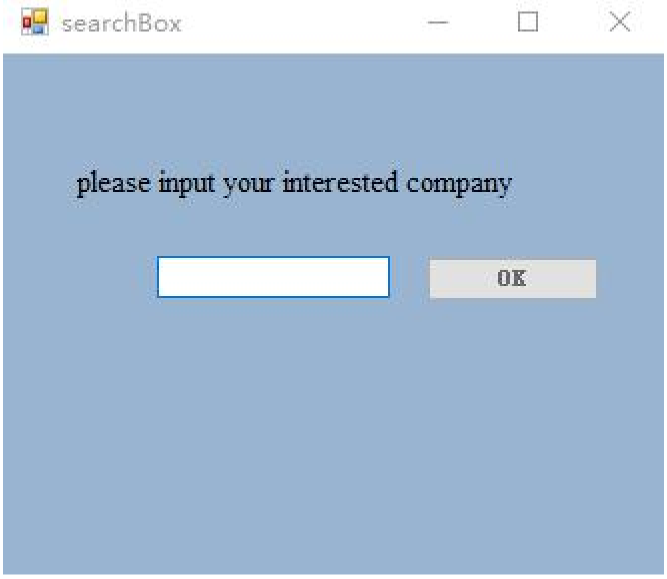


Fig 1: User input form

- **Link Generation:**
  Once the user has provided the input, relevant links are generated that will be used to extract the data. These links are generated based on the type of information that will be displayed to the user.

- **Link Database:**
  Once the links are generated from the above step. These links are then stored in a SQL database. The different fields of the link database are shown in the figure below.

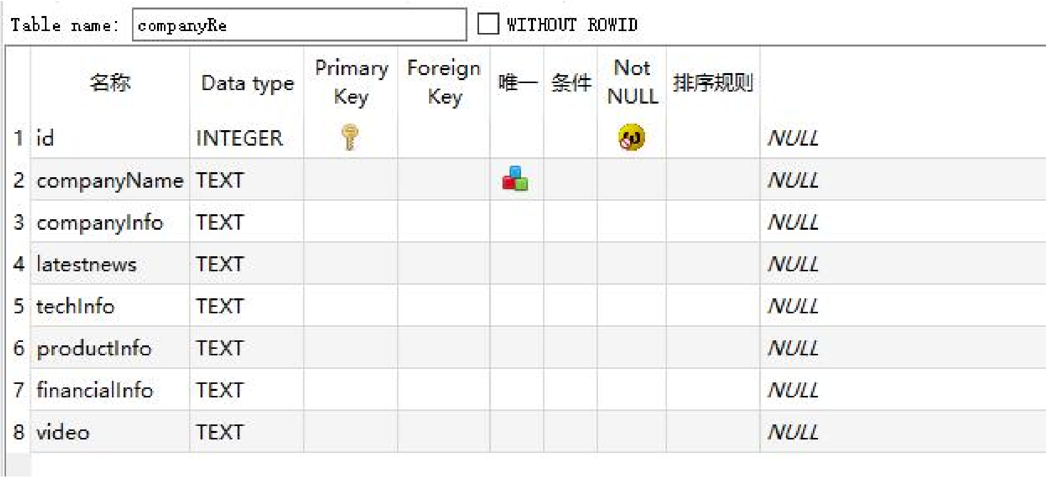


Table name: companyRe □ WITHOUT ROWID

	名称	Data type	Primary Key	Foreign Key	唯一	条件	Not NULL	排序规则	
1	id	INTEGER	✓				✓		*NULL*
2	companyName	TEXT			✓				*NULL*
3	companyInfo	TEXT							*NULL*
4	latestnews	TEXT							*NULL*
5	techInfo	TEXT							*NULL*
6	productInfo	TEXT							*NULL*
7	financialInfo	TEXT							*NULL*
8	video	TEXT							*NULL*

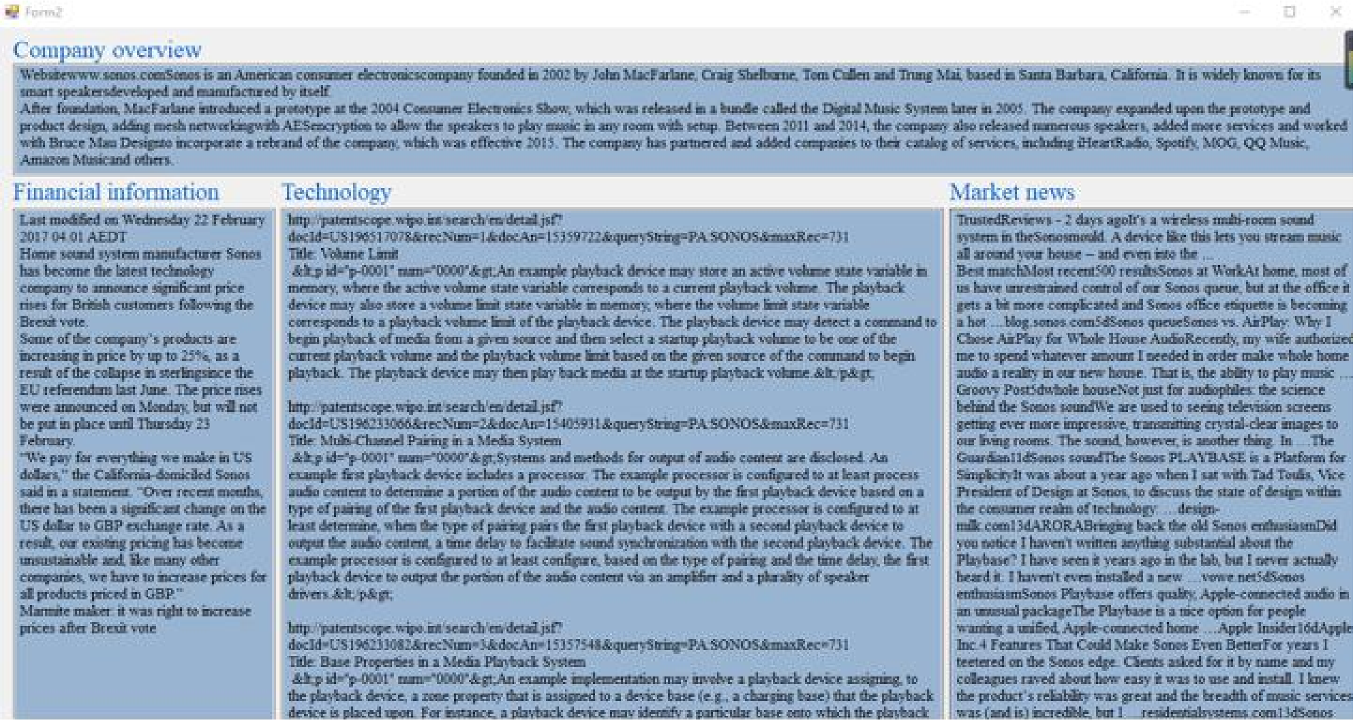
Fig 2: Link database fields

- **Web Crawling and Data Extraction:**
  Once we have stored all the links that will be used to obtain information, the next will be to extract data from these links and filter the content, so that only the relevant information is displayed to the user.

- **Information Database:**
  The data that is extracted from the above step is then stored in another database. The fields of the database and their description is provided in the table below:

Name of Field	Description
id	Works as the primary key for the database.
companyName	Stores the name of the company which was input by the user in the search input form.
companyInfo	A brief introductory information about the company is stored.
latestNews	The relevant latest news about the company is stored in this field.
techInfo	Technical information about the patents of the company are stored.
productInfo	Information on the products that the company manufactures are stored in this field.
financialInfo	This field stores the financial information about the company such as its financial reports, revenue, net profit and stock market value.
video	This field contains a link to an introductory video about the company or the product.

- **Output:**
  The output is displayed in the form of a windows form application which basically retrieves the information from the database and displays it to the user.



Company overview

Websitewww.sonos.comSonos is an American consumer electronicscompany founded in 2002 by John MacFarlane, Craig Shelburne, Tom Cullen and Trung Mai, based in Santa Barbara, California. It is widely known for its smart speakersdeveloped and manufactured by itself.
After foundation, MacFarlane introduced a prototype at the 2004 Consumer Electronics Show, which was released in a bundle called the Digital Music System later in 2005. The company expanded upon the prototype and product design, adding mesh networkingwith AESencryption to allow the speakers to play music in any room with setup. Between 2011 and 2014, the company also released numerous speakers, added more services and worked with Bruce Mau Designto incorporate a rebrand of the company, which was effective 2015. The company has partnered and added companies to their catalog of services, including iHeartRadio, Spotify, MOG, QQ Music, Amazon Musicand others.

Financial information

Last modified on Wednesday 22 February 2017 04.01 AEDT
Home sound system manufacturer Sonos has become the latest technology company to announce significant price rises for British customers following the Brexit vote.
Some of the company's products are increasing in price by up to 25%, as a result of the collapse in sterlingsince the EU referendum last June. The price rises were announced on Monday, but will not be put in place until Thursday 23 February.
"We pay for everything we make in US dollars," the California-domiciled Sonos said in a statement. "Over recent months, there has been a significant change on the US dollar to GBP exchange rate. As a result, our existing pricing has become unsustainable and, like many other companies, we have to increase prices for all products priced in GBP."
Marmite maker: it was right to increase prices after Brexit vote

Technology

http://patentscope.wipo.int/search/en/detail.jsf?docId=US196517078&recNum=1&docAn=15359722&queryString=PA:SONOS&maxRec=731
Title: Volume Limit
<p id="p-0001" num="0000">An example playback device may store an active volume state variable in memory, where the active volume state variable corresponds to a current playback volume. The playback device may also store a volume limit state variable in memory, where the volume limit state variable corresponds to a playback volume limit of the playback device. The playback device may detect a command to begin playback of media from a given source and then select a startup playback volume to be one of the current playback volume and the playback volume limit based on the given source of the command to begin playback. The playback device may then play back media at the startup playback volume.</p>

http://patentscope.wipo.int/search/en/detail.jsf?docId=US196233066&recNum=2&docAn=15405931&queryString=PA:SONOS&maxRec=731
Title: Multi-Channel Pairing in a Media System
<p id="p-0001" num="0000">Systems and methods for output of audio content are disclosed. An example first playback device includes a processor. The example processor is configured to at least process audio content to determine a portion of the audio content to be output by the first playback device based on a type of pairing of the first playback device and the audio content. The example processor is configured to at least determine, when the type of pairing pairs the first playback device with a second playback device to output the audio content, a time delay to facilitate sound synchronization with the second playback device. The example processor is configured to at least configure, based on the type of pairing and the time delay, the first playback device to output the portion of the audio content via an amplifier and a plurality of speaker drivers.</p>

http://patentscope.wipo.int/search/en/detail.jsf?docId=US196233082&recNum=3&docAn=15357548&queryString=PA:SONOS&maxRec=731
Title: Base Properties in a Media Playback System
<p id="p-0001" num="0000">An example implementation may involve a playback device assigning, to the playback device, a zone property that is assigned to a device base (e.g., a charging base) that the playback device is placed upon. For instance, a playback device may identify a particular base onto which the playback

Market news

TrustedReviews - 2 days agoIt's a wireless multi-room sound system in theSonosmould. A device like this lets you stream music all around your house – and even into the ...
Best matchMost recent500 resultsSonos at WorkAt home, most of us have unrestrained control of our Sonos queue, but at the office it gets a bit more complicated and Sonos office etiquette is becoming a hot ...blog.sonos.com5dSonos queueSonos vs. AirPlay: Why I Chose AirPlay for Whole House AudioRecently, my wife authorized me to spend whatever amount I needed in order make whole home audio a reality in our new house. That is, the ability to play music ... Groovy Post5dwhole houseNot just for audiophiles: the science behind the Sonos soundWe are used to seeing television screens getting ever more impressive, transmitting crystal-clear images to our living rooms. The sound, however, is another thing. In ...The Guardian11dSonos soundThe Sonos PLAYBASE is a Platform for SimplicityIt was about a year ago when I sat with Tad Toulis, Vice President of Design at Sonos, to discuss the state of design within the consumer realm of technology. ... design-milk.com13dARORABringing back the old Sonos enthusiasmDid you notice I haven't written anything substantial about the Playbase? I have seen it years ago in the lab, but I never actually heard it. I haven't even installed a new ... vowe.net5dSonos enthusiasmSonos Playbase offers quality, Apple-connected audio in an unusual packageThe Playbase is a nice option for people wanting a unified, Apple-connected home ...Apple Insider16dApple Inc.4 Features That Could Make Sonos Even BetterFor years I teetered on the Sonos edge. Clients asked for it by name and my colleagues raved about how easy it was to use and install. I knew the product's reliability was great and the breadth of music services was (and is) incredible, but I ... residentialsystems.com13dSonos

Fig 3: Shows the output provided to the user

Technical Approach

According to the product or company information submitted by the end-user, web crawlers go to several data sources, bring back the top pages individually. For different product categories and different industries, different hubs, authority pages, search engines are selected to be data sources, company's home page and related pages are also included into the upcoming project's data sources.

Then these top pages are content filtered (irrelevant data is deleted), stored into the database, indexing. The data is fetched from the database according to the end-user's submitted information, create operable page elements to contain each kind of data, for example, technical data, financial data and the latest marketing news.

Functions units:

- Input and Output forms: retrieves the user's input and displays the output to the user
- Web crawlers: go to the link generated to retrieve relevant information, bring back top webpage contents.
- Content filter: filter irrelevant information, like html tags, advertising.
- Database: fields include industry's category, website data sources, company's name, product' name, technical data, financial data, news data etc.

Resources:

Resources	Details
Project team	A team of two members working on the project.
Computers	Team members will use their personal computers for the task of development.
Programming Language	C#
IDE	Microsoft Visual Studio Community 2017
Code Management Tools	GitHub
Database Tools	MySQL, SQLite Studio
External Libraries	Lucene.net, HTML Agility Pack

Results

Our system provides an interface for the user to input the name of the company they wish to obtain information about. The output displays the all the relevant information to the user in the form of a windows form. The information provided by the system includes introductory information about the company, latest company specific news, technical information about some of the company's patents, information about various products produced by that company and an overview of the financial situation of the company.

Discussion and Future work

The application we have developed is very basic and many features can be added and improved. But given the timeframe only a basic idea took shape but this idea has the potential to be developed into an application that can be useful to millions of people that wish to obtain information about various companies. Some of the features that can be added to our system in the future are:

- The search fields are increased so that user can input not only the company name but can select from an image list of the top most searched companies.
- The search field can also have different product categories to choose from.
- The content retrieved also needs to be filtered in such a way that the most useful information is extracted from the webpage.
- The financial information section of the output can have an addition feature that displays different charts and tables to give user a better financial perspective.

Reflection on Project Presentation

Presentation skills are very important as they require organisation and content. Presentations skills are vital to convey your idea to fellow colleagues, potential investors and stakeholders. An idea that one possesses can be of significant importance but has no relevance until and unless it can be conveyed to others. The purpose of a presentation is not to convey information. You can do that with a data sheet. The purpose of a presentation is to show how that information has meaning for whatever is going on in the businesses and lives of the members of your audience [3].

There are many kinds of presentation which can range from a couple of minutes to a detailed half an hour one. A presentation that only lasts a few minutes should give a very high-level description of the work and should convey the message to an audience that has very limited understanding of the field.

Some of the strategies that one should incorporate to prepare and deliver a good presentation are [4]:

- The slides should contain less text and more images
- The font size should be big
- The use of illustrations to get across the key concept
- Use of keywords rather than having long sentences
- Making eye contact with the audience
- Use of body language to convey the message
- The presentation should be presented as a story

References

[1] "Techcrunch".https://en.wikipedia.org/wiki/TechCrunch#CrunchBase . N.p., 2017. Web. 29 May 2017.

[2] "Owler".https://en.wikipedia.org/wiki/Owler. N.p., 2017. Web. 29 May 2017.

[3] "5 Essential Rules For Great Presentations".https://www.inc.com/geoffrey-james/5-essential-rules-for-great-presentations.html. N.p., 2017. Web. 29 May 2017.

[4] "The Importance Of Presentation Skills In The Workplace - Leaders In Heels".https://leadersinheels.com/career/public-speaking/importance-presentation-skills-workplace/. N.p., 2017. Web. 29 May 2017.

[5] "Are There Any Other Sites Like Crunchbase".https://www.quora.com/Are-there-any-other-sites-like-Crunchbase-com. N.p., 2017. Web. 29 May 2017.

README

The application called WebCrawlertest, which is used to search company’s data, including company overview, financial information, latest market news and technical information.

The application is developed by C# , Microsoft Visual Studio 2015 and SQLiteStudio.

The database is test.db locate in...\WebCrawlertest\WebCrawlertest\bin\Debug. To use this database, System.Data.SQLite.dll and System.Data.SQLite.Linq.dll should be added to the project. There are two tables in the test.db, one is companyRe, the other is searchResults, companyRe table is used to collect relevant links of the company, and searchResults table is used to store the extracted content from relevant links of the company.

One piece of test data record is stored to the database, wherein, the technical information’s link is a RSS link, and the technical content stored in searchResults is the content extracted by the method

public static string wipoRSSItem(string rss).

Other links stored in companyRe are HTML links. Html2Article.dll is used to process them. Html2Article.dll is a tool found online which can extract the main body from the whole HTML file. Therefore, to claim using StanSoft; to add the Html2Article.dll to the project.

Form1.cs is used to receive the user’s input, and Form2.cs is used to display the content fetched from searchResults table of test.db.

When running the program, the Form1 will pop up first. After end-user inputs interested company name and click the button OK, the Form1 will disappear. The Form2 will pop up to display the company’s key information(such as company overview, financial information, market news and technical information)

There is only one piece of test data in database CompanyRe, testers can input **Sonos(a company)** to test the application.