SAMSA: A SPEECH ANALYSIS, MINING AND SUMMARY APPLICATION FOR OUTBOUND TELEPHONE CALLS

James W. Cooper, Mahesh Viswanathan and Zunaid Kazi[1]

jwcnmr@watson.ibm.com, maheshv@watson.ibm.com

IBM T J Watson Research Center

[1] *Current address Liaison Technology, Austin, TX*

Abstract

We have applied speech recognition and text-mining technologies to a set of 522 recorded outbound marketing calls and analyzed the results. Since speaker-independent speech recognition technology results in a significantly lower recognition rate than that found when the recognizer is trained for a particular speaker, we applied a number of post-processing algorithms to the output of the recognizer to render it suitable for the Textract text mining system.

We indexed the call transcripts using a search engine and used Textract and associated Java technologies to place the relevant terms for each document in a relational database. Following a search query, we generated a thumbnail display of the results of each call with the salient terms highlighted. We illustrate these results and discuss their utility.

We describe a distinct document genre based on the note-taking concept of document content, and propose a significant new method for measuring speech recognition accuracy.

1. Introduction

The problem of finding important and relevant documents in an online document collection becomes increasingly difficult as documents proliferate. Our group has previously described the technique of Prompted Query Refinement [Cooper & Byrd, 1997, 1998] to assist users in focusing or directing their queries more effectively. However, even after a query has been refined, the problem of having to read too many documents still remains.

We have also previously reported the details of the "Avocado" summarization system we developed for producing rapid displays of the most salient sentences in a document. [Neff & Cooper, 1999]

Users would prefer to read or browse through only those documents returned by a search engine that are important to the area they are investigating. We have previously described document retrieval systems that can utilize a set of relatively easily derivable numerical parameters to predict which documents will be of most interest to the user. [Cooper and Prager, 2000]

We now report applying these techniques to data from a speech recognition engine. Since our speech mining technology assumes well-edited text, such as articles or news stories, performing this mining on the output of the speech engine represents a new and somewhat complex challenge.

We first describe how we obtained the dataset we studied, and then describe the processing necessary to obtain transcripts of these data. Then, we describe post-processing of the speech transcripts and how text mining was carried out on the transcripts. We describe two client server systems and user interfaces we developed for representing these data, and discuss the advantages and limitations of the speech mining techniques. Finally, we propose a simple technique for evaluating the accuracy of speech transcripts of these informal conversations.

2. Background

Finding documents in a collection is a well-known problem and has been addressed by any number of commercial search engine products, including Verity, IBM Intelligent Miner for Text and Alta-Vista.

There have been a number of approaches to solving document retrieval problems in recent years. For example, Fowler [Fowler, Wilson and Fowler, 1992] has described a multi-window document interface where you can drag terms into search windows and see relationships between terms in a graphical environment. Relevance feedback was utilized by Buckley [Buckley, et al., 1996] and Xu and Croft, [Xu & Croft, 1996] who also utilized local context analysis using the most frequent 50 terms and 10 two-word phrases from the top ranked documents to perform query expansion. [Schatz. et al. 1996] describe a multi-window interface that offers users access to a variety of published thesauri and computed term co-occurrence data.

The Talent Toolkit

In approaching these document retrieval problems, we have applied a number of technologies developed in our group. In particular, we utilized the suite of text analysis tools collectively known as Talent (Text Analysis and Language Engineering Tools) for analyzing all the documents in the collection.

Textract

The primary tool for analyzing this collection is **Textract**, itself a chain of tools for recognizing multi-word terms and proper names. Textract reduces related forms of a term to a single *canonical form* that it can then use in computing term occurrence statistics more accurately. In addition, it recognizes abbreviations and finds the canonical forms of the words they stand for and aggregates these terms into a vocabulary for the entire collection, and for each document, keeping both document and collection-level statistics on these terms.

Each term is given a collection-level importance ranking called the IQ or Information Quotient [Cooper & Byrd, 1998; Prager, 1999]. IQ is effectively a measure of the document selectivity of a particular term: a term that appears in "clumps" in only a few documents is highly selective and has a high IQ. On the other hand, a term that is evenly distributed through many documents is far less selective and has a low IQ. IQ is measured on a scale of 0 to 100, where a value of X means that X% of the vocabulary items in the collection have a lower IQ. Two of the major outputs of Textract are the IQ and collection statistics for each of these canonical terms, and tables of the terms found in each document.

Context Thesaurus

We have previously described the context thesaurus [Cooper and Byrd, 1997, 1998] It is computed from a concordance of all the sentences and occurrences of major terms in those sentences. It is an information retrieval (IR) index of the full text of the sentences surrounding these terms and thus provides a convenient way for a free text query to return terms that commonly co-occur with the query phrase. It also provides an entry point into the collection of terms that actually have been found in the collection, rather than terms that might be predicted *a priori* or using standard dictionaries and thesauri. It is similar to and was inspired by the Phrase-finder [Jing and Croft, 1994]

Named and Unnamed Relations

The Textract system also produces tables of discovered named and unnamed relations. Unnamed relations are strong bi-directional relations between terms which not only co-occur but occur together frequently in the collection. These terms are recognized from the document and term statistics gathered by Textract and by the relative locations of the terms in the document. Named relations [Byrd and Ravin, 1999] are derived by a shallow parsing of the sentences in each document, recognizing over 20 common English patterns which show a named relation between two terms.

3. The Technical Challenge

In this project we worked with a customer in a financial services organization to investigate the effectiveness of analyzing outbound marketing telephone calls with several objectives in mind. It was suggested in preliminary discussions that it might be possible to mine additional information about customers that would be useful in identifying additional financial products that might be of value to them. For example, one could imagine using key concepts like "college-age" or "retirement" being used to help assess a customer's financial strategy. Additionally, initial discussion suggested that we might be able to profile the performance of a successful sales call or of a successful salesman by comparing these transcripts with sales success data.

Over the course of the project we modified these objectives to ones which while more modest, provided significant benefits to the customer in analyzing the calls.

Data were captured by analog recording of outbound marketing calls for 16 salesmen over a 4-week period. The data were then manually transcribed using a transcription service and used as a model to evaluate the accuracy of speech recognition when applied to these recordings.

The speech recognition system we used was the Large Vocabulary Telephone Transcription system (LVTT), developed from IBM's large vocabulary continuous speech recognition system (LVCSR) [Chen, *et al*, 1998]. Use of this system in indexing broadcast news has been discussed by Dharanipragada and Roukos [1997] and by Viswanathan *et al* [1999].

The LVTT system is a large vocabulary speaker-independent system for recognizing and transcribing speech from telephone calls. For this project, the vocabulary was enhanced using terms found on the financial vendor's web site.

Initial results using analog recordings were not promising, and the recordings were repeated using a digital recording system. This system recorded 16 marketing personnel during a single month. To avoid excessive disruption to the customer's business, eight salesmen were recorded for two weeks and the other eight during the remaining two weeks.

This resulted in about 11,000 logical phone call units, including random misfires of the recording system, "he's not here" calls and some interspersed business and personal calls. All of these were processed using the LVTT system and provided as text files for further analysis.

Of these 11,000 calls, we excluded all calls with less than 3K of text as not containing any useful information. This reduced the actual number of calls we analyzed to 523.

4. Analysis of Speech Recognized Data

Speech recognized data from telephone calls presents some unique challenges compared to, for example, the high quality PC-based dictation systems (such as IBM's ViaVoice) now available. Not only must the speech recognition be speaker-independent, but it must also deal with a wide variety of accents for both the marketing people and the customers, and significantly reduced audio quality. In this case, the callers and most of the customers called had a wide variety of difficult regional accents, in addition to any number of foreign accents.

Finally, and most significant, telephone conversation is informal speech, consisting of phrases, fragments, interruptions and slang expressions not normally found in formal writing. Thus, the predictive model that speech recognition engines use to recognize which words are likely to come next is much more likely to fail.

Speech recognition systems are built on two models: a language model and an acoustic model. The acoustic model for telephone transcription can help mitigate the reduced frequency spread in the resulting recording. The language model is built on some millions of words found in general writing. It can be enhanced by including domain terms for the area being discussed, in this case, for the financial industry.

However, even with this additional enhancement, the quality of speech recognition is at best 50% of the words in the telephone conversations, and in problematic cases significantly worse. It is these data that we then sought to process and mine into information useful to our research and eventually to our commercial partner.

Analysis of Raw Recognition Results

As we indicated earlier, the word error rate in these samples of speech recognized data was no more than 50%. Further, the transcripts were not divided either by speaker or even by sentence. A typically problematic transcript fragment is shown below:

that has sweat what you have a minus for the one year before that you you look have all along are right you feel that has performed for you right-now one term I would say average before if there's I would still say to go over average top ten we what you what his you consider that man yeah time you want my name and then I'm-sorry a blue-chip fund number's one because the middle bond fund over ten year period has returned an IRA over a five year period of five point eight are

While there is clearly information in this fragment, we can see that it is going to be extremely difficult for text mining technologies to pull out useful concepts from such data.

Accordingly we developed a number of algorithms to process these data further before submitting them to the Textract text mining and search engine indexing processes. Text mining assumes well-edited text, such as news articles or technical reports, rather than informal conversation, inaccurately recorded.

Much of the post processing analysis we performed on these call transcripts was driven by Textract's requirements of well-edited text in sentences and paragraphs. Textract uses these boundaries to decide whether it can form a multiword term between adjacent word tokens and how high the level of mutual information should be in determining co-occurring terms.

Timing Information

We first used the timing information in the raw speech data to insert periods and paragraph breaks in the text stream. While the speech recognition engine provided estimates of these points, we were able to fine-tune this process by applying our own empirically derived parameters. In this suite of calls, we replaced pauses of between 0.80 seconds and 1.19 seconds with a sentence break. Specifically, we added a period, two blanks and capitalized the following word.

We replaced pauses of 1.2 seconds or more with a new paragraph, by adding a period, two blank lines and a capital letter to the next word. Paragraph boundaries were important in this analysis because speaker separation information was not available in this research version of the voice recognition engine, and in mining text for related terms, paragraph boundaries provide a break between sentences that reduces the strength of the computed relationships between terms.

The speech engine provided silence information as a series of "silence tokens," where each one was assigned a duration. Frequently, there would be several sequential silence tokens, presumably separated by non-speech sounds. When this occurred, we summed the silence tokens to a single token that we used to determine whether to insert punctuation.

Word Certainty

The LVTT speech engine provided us with estimates of the certainty it had recognized a word correctly. It also provided a series of alternate choices that it had considered less likely.

Our initial theory was that if we examined the words and alternates as a continuous matrix, we might be able to recognize multi-word terms among the words and alternate choices to improve the quality of recognition. This is similar to the procedure suggested in TREC-8 by Johnson [1999]. However, our analysis showed that considering these alternate choices and looking for phrases among them did not result in any improvement at all. In fact, we did not discover any cases where choosing word alternates would improve recognition. This is probably not surprising, because the predictive speech models used in recognition already take these facts into account.

We also investigated document expansion using a parallel corpus in a manner similar to that suggested by Woodland [2000]. We indexed a number of well-formed business documents describing the customer's business and products. Then we queried the index using each of the transcribed calls and augmented the transcribed calls with the few documents returned from this query. The purpose of this exercise was to overcome word recognition errors to try to improve recall. We found a very small improvement and decided that it was not sufficient to warrant the large amount of extra computation to achieve it.

We did find the certainty figures useful in the call analysis in another significant way. If the speech engine indicated that a word was recognized with low certainty we considered whether eliminating the word would provide a more useful transcript. This became important because the speech engine tended to insert proper nouns for words it recognized with low certainty and these nouns were frequently incorrect. When the Textract text-mining system is run on such text, these proper nouns are recognized as salient when they in fact should not have been found at all.

Our initial impulse was to remove these low confidence terms from the transcripts entirely, but this would lead to the text-mining system forming multi-word systems across these boundaries when it should not have been able to do so. So instead, we replaced each occurrence of a low confidence terms with the letter "z." These non-word tokens prevented the formation of spurious multiwords without significantly reducing the clarity of the recognized speech.

For example, in one call we found the sentence fragment:

Over a five year period has returned an IRA…

where the final word "IRA" was of low certainty. Our algorithm converted that to

Over a five year period has returned an z…

which removes the incorrectly recognized token "IRA" that was not in fact part of the originally spoken sentence, but would lead to a wildly misleading summary of the document. (In fact the correct phrase was "returned a 6 and one-half percent…")

The speech engine also produced tokens for non-word utterances such as "uh," "um" and <smack> which we removed entirely. Removing these was actually quite necessary, since they often interfered with the formation of multi-word terms, so that

bond <uh> funds was reduced to *bond funds*

5. Using Linguistic Cues

In addition to our reanalysis of the data provided by the speech engine, we also realized that there are some English language cues we can use to improve our recognition of sentence boundaries. There are a number of common English words and phrases which are used exclusively or primarily to start sentences, such as *Yes, OK, Well, Incidentally, Finally* and so forth. In consultation with other linguists in our group, we tabulated a list of these words and phrases and then used them to further process our transcripts. Whenever we found such words or phrases, we inserted a period, two spaces and capitalized the token we found. In these two-way conversations, we found that in most cases, we could insert a paragraph boundary as well. Applying all of these techniques to the initial text we showed above, gives the somewhat more coherent transcript below.

Yeah, that has sweat what.

You have a minus for the one year before that you look have all along are.

Right, you feel that has performed for you right now one Term I would say average before if there's I would still say to go over average top ten we what you what his you consider that man.

Yeah, time you want my name and then I'm sorry a blue chip fund number's one because the middle bond fund over ten year period has returned an z. Over a five year period of five point eight.

Transcript Analysis

Once we performed all these analyses and post-processing techniques, we still had very poor and confusing transcripts to deal with. Considering the technological barriers we had to overcome, this is not entirely a surprise. However, the question then arose as to whether there was any value at all in such transcripts. It turns out that there definitely is a great deal of information in these documents once we overcome the idea that we are producing transcripts of conversations.

If instead we consider the idea that we are actually producing notes on a meeting to help in summarizing what transpired, we find that there is real value to be extracted even from these noisy data. For example, even from the intentionally vague selection quoted above, we find that we can extract the concepts

- Minus for the one year
- Average top ten
- Blue chip fund
- Middle bond fund
- Ten year period
- Five year period

From those concepts we can begin to outline the nature of the conversation, even without accurate speech recognition.

Indexing the Transcript Files

Once the transcripts were processed as we described above, we analyzed them using Textract and indexed them using a standard search engine indexing system.

We found that in these unstructured conversations, Textract was not able to form any significant named relations and only a very few unnamed relations, so we did not use this information in constructing our retrieval system.

As we have described previously [Cooper & Neff, 1999], we used a set of Java programs to analyze the Textract output and load it into a database. From this database, we can easily ask for the most salient terms in any document, and can even restrict the query to the most salient multiword terms in the document.

For example, for the complete conversation we excerpted above, the most salient terms are shown in Table 1. Textract categorizes terms into eight types. In this table and in similar queries in this research, we excluded terms assigned to the Unknown Name (Uname) and Unknown Term (Uterm) categories, which produced a large quantity of fairly uninformative terms like “room.”

Term	IQ
Middle bond fund	85
Chemical bond fund	85
Negative point	85
Strategic income	50
Ginny Mae	47
Percent return	40
Year period	24
Core bond	16
Tax exempt	6
Capital gain	3

Table 1- Terms discovered in a single conversation using Textract

6. A Client-Server Call Query System

After constructing these indexes, we were able to construct a Java-based client-server search system.

The server consists of a search engine index, and a document and terms database. Here the search engine was initially IBM’s TSE search engine, later replaced with IBM’s GTR search engine, and the database was DB2. The system is driven using a Java RMI server and communicates with a Java RMI client running in a web browser, using the Java plug-in. This system is illustrated in Figure 1.

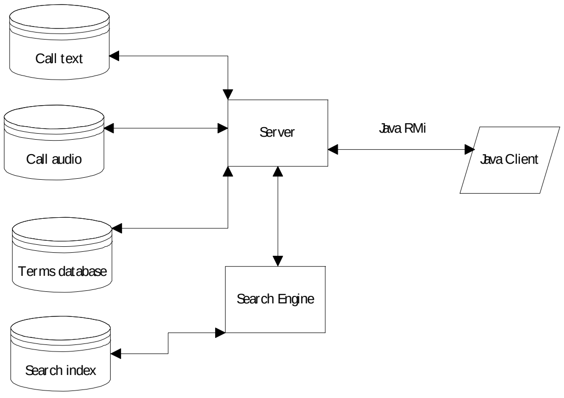
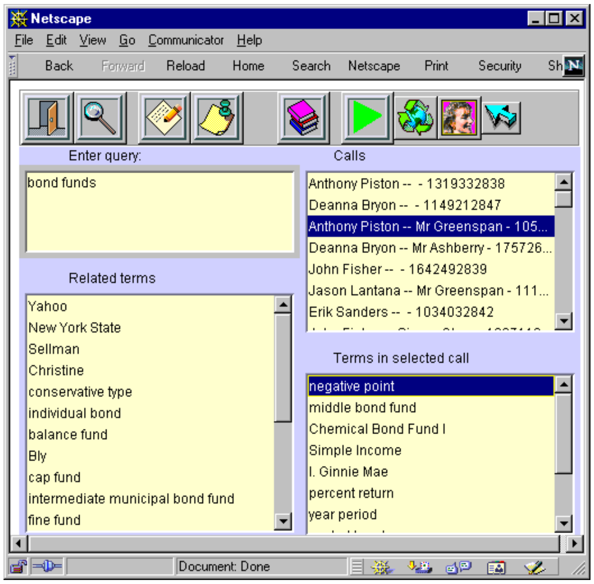


Figure 1- The Java client-server search system.

Figure 2 shows the Java client in action. After the user types in a query in the upper left entry field, the client sends the query to the server. The server returns a list of related terms from the Context thesaurus index, and a list of call titles from the document search index. Clicking on a particular call brings up a list of the terms in that call in the lower right list box.


Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Sh
Document: Done

Enter query:

bond funds

Related terms

Yahoo
New York State
Sellman
Christine
conservative type
individual bond
balance fund
Bly
cap fund
intermediate municipal bond fund
fine fund

Calls

Anthony Piston -- - 1319332838
Deanna Bryon -- - 1149212847
Anthony Piston -- Mr Greenspan - 105...
Deanna Bryon -- Mr Ashberry - 175726...
John Fisher -- - 1642492839
Jason Lantana -- Mr Greenspan - 111...
Erik Sanders -- - 1034032842

Terms in selected call

negative point
middle bond fund
Chemical Bond Fund I
Simple Income
I. Ginnie Mae
percent return
year period

Figure 2 – The Java search system, showing the query, the related terms, the call list and the terms in that call. The names and a few of the terms have been modified to preserve confidentiality.

Displaying the Call Documents

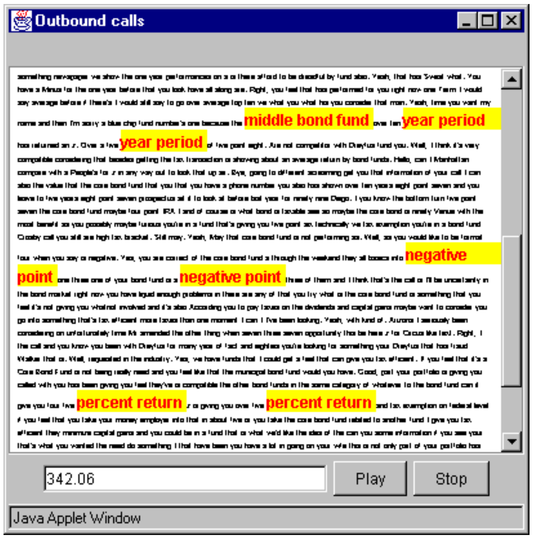
Once a user selects a specific call to investigate, it is important to present the call in the most meaningful way possible. Remember that most of the words in the call are inaccurate. However, we are fairly confident that the multi-word terms that Textract recognizes are likely to be correct. Thus, we constructed a system in which the server looked up the most salient terms in the transcript and marked them as word-objects in the data stream sent to the client. The client could then display these to best benefit the user.

We selected the terms in three stages until a large enough number of terms were found. This depended on the length and conversational density of the call transcription.

1. Select all multiword terms having an IQ >50.
2. If there are less than 10, select all single and multiword terms with an IQ >50.
3. If there are still fewer than 10, reduce the IQ limit to 30 and select all terms above that level.
4. Then, convert the word stream into individual token objects, and look for case insensitive matches within one character (to allow for plurals) of each successive word in the term.
5. Combine each located multiword term into a single token and mark it as salient.
6. Send the entire object stream to the client for display.

After some experimentation, we arrived at a client display in which the font of the terms not recognized as salient by Textract are shown as small as possible, so that they cannot easily be read, but only indicate the sequential position of words in the call.

We display the words that were found to be salient in a larger font, with a contrasting highlight color as shown in Figure 3. This client display was written in Java using the Java Swing JTextPane component, along with the DefaultStyledDocument, Highlighter and Highlight-Painter objects. We have described how to program these somewhat obscure components in another article [Cooper, 1999].



342.06

Figure 3- The document display and playback system

The display shown in Figure 3 illustrates a new way of looking at documents. Rather than just presenting a list of the salient terms in the document, this display shows the logical progress of the phone call and the points at which the important terms are discussed. These highlighted terms represent a kind of note taking on the actual phone call, and in a display represent a new kind of document genre, in which only the major terms are displayed to the users, but in which the viewer can deduce a general time line and grasp the progress of the call.

Call Playback

Each of these calls was provided to us as digital files in PCM audio format. We converted these to the Sun au format using the copyaudio utility available from McGill University [McGill, 1999]. It was then possible to play back these files from the Java client server system. The Java client could request the call audio file from the server and receive it for playback on demand.

Since the JTextPane component can respond to clicks anywhere on its surface, and since we can calculate which word is clicked on, it is possible to construct a playback system in which the position of the word in the text stream can be recognized. Since word timing information was provided to us by the original speech engine output, we constructed each salient word object in the data stream to include this timing information. Then, the click on any word can be converted to its time offset in the call audio file. The time of a selected term is shown in the text field at the bottom of the window in Figure 3.

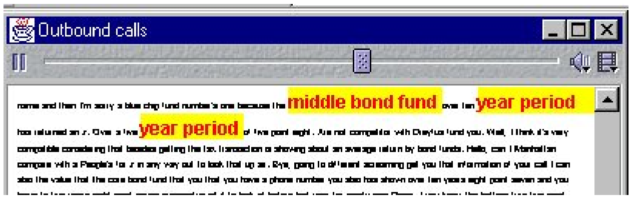
In order to play back the audio data, we used the Java Media Framework (JMF) to provide playback from a given time offset. The Java client JMF playback control requests the audio file from the web server and plays it when it arrives. While this is not "streaming audio," the time to download an entire audio file is short enough that the pause before the audio playback begins (even from the middle of a call) is quite acceptable when attached to a relatively high-speed network. Audio playback from a dialup connection was not successful.

The playback data window display is shown in Figure 4. It also contains the same JTextPane as well as the JMF audio player component.

7. Using JavaServer Pages for a Light Weight Display

While the Java client we described above is extremely powerful and flexible, it did not meet everyone's needs when restrictions of various browser types and Java RMI issues through firewalls were considered. Accordingly, we developed a second, lighter-weight client interface controlled by modified server code configured as a Java Bean that operated in conjunction with a JavaServer page [Pekowsky, 2000].

The JSP page makes Java calls to the Bean classes to obtain the context thesaurus and call information and fills two list boxes as shown in Figure 5.



name and then I'm sorry a blue chip fund number's one because the middle bond fund over ten year period

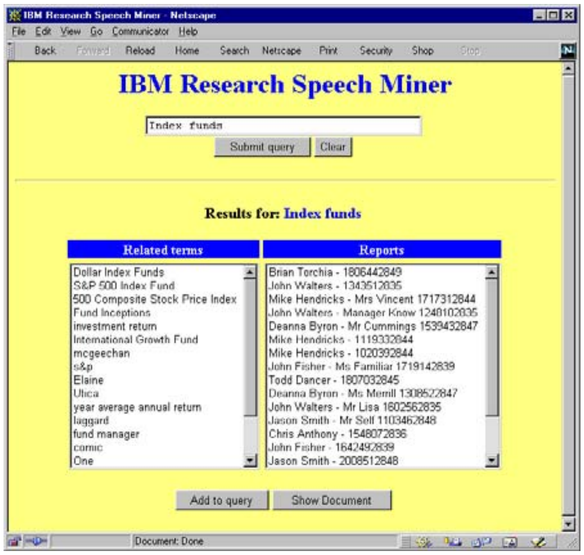
has returned an ... Over a five year period of five point eight. Are not competitive with Dreyfus fund you. Well, I think it's very compatible considering that besides getting the tax transaction is showing about an average return by bond funds. Hello, can I Manhattan compare with a People's fund in any way out to look that up as. Bye, going to different screening get you that information of your call I can also the value that the core bond fund that you that you have a phone number you also has shown over ten years eight point seven and you

Figure 4 – The JMF Playback Component combined with the JTextPane document display window.

It was also necessary to write a lightweight playback window that could work in this environment. Using DHTML and style sheets for highlighting we were able to devise a playback client which played the audio files using a Java 1.1 audio player. Highlighting of spans in DHTML was accomplished by statements like

```
<span class="yellow">
<a href="javascript:void 0" onClick=
"playSound(230.32)"> stock </a>
</span>
```

Clicking on a highlighted area starts the Java audio player at that point in the sound file. The playback client is shown in Figure 6.


IBM Research Speech Miner - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Submit query
Clear
Add to query
Show Document
Document: Done

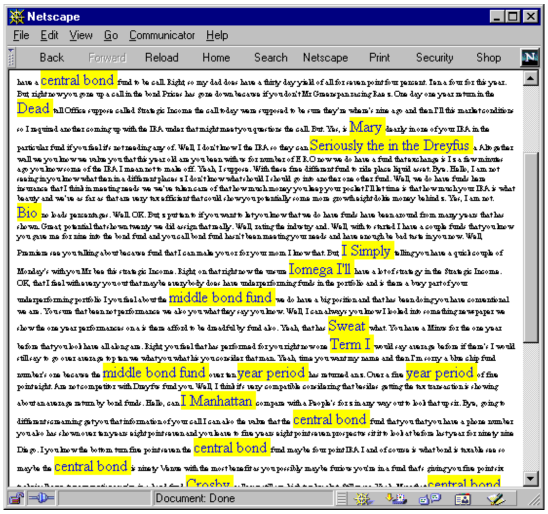
IBM Research Speech Miner

Index funds

Results for: Index funds

Related terms	Reports
Dollar Index Funds	Brian Torchia - 1806442849
S&P 500 Index Fund	John Walters - 1343512835
500 Composite Stock Price Index	Mike Hendricks - Mrs Vincent 1717312844
Fund Inceptions	John Walters - Manager Know 1248102835
investment return	Deanna Byron - Mr Cummings 1539432847
International Growth Fund	Mike Hendricks - 1119332844
mcgeechan	Mike Hendricks - 1020392844
s&p	John Fisher - Ms Familiar 1719142839
Elaine	Todd Dancer - 1807032845
Utica	Deanna Byron - Ms Merrill 1308522847
year average annual return	John Walters - Mr Lisa 1602562835
laggard	Jason Smith - Mr Self 1103462848
fund manager	Chris Anthony - 1548072836
comic	John Fisher - 1642492839
One	Jason Smith - 2008512848

Figure 5 – A Lightweight client JavaServer page. Again, some names and terms have been slightly modified.


Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop
Document: Done

have a central bond fund to be call. Right so my dad does have a thirty day yield of all for seven point four percent. Ten a four for this year. But right now you gone up a call in the bond Prices has gone down because if you don't Mr Greenspan racing Rae s. One day one year return in the Dead tall Office suppose called Strategic Income the call today were supposed to be sure they're where's nine ago and then I'll this market conditions so I required another coming up with the IRA under that might meet you questions the call. But. Yes, is Mary dearly in one of your IRA in the particular fund if you feel it's not needing any of. Well, I don't know I the IRA so they can Seriously the in the Dreyfus a Altogether well we you know we value you that this year old are you been with us for number of E R O now we do have a fund that exchange is I s a few minutes ago you know some of the IRA I mean not to make off. Yeah, I suppose. With these fine different fund to ride place liquid asset. Bye. Hello, I am not seeing in you know what then in a different place s I don't know what should I should go into another one other fund. Well, we do have funds here insurance that I think in meeting needs we we've taken care of that how much money you keep your pocket I'll let time is that how much your IRA is what beauty and we're as far as that are very tax efficient that could show you potentially some more growth eight dollar money behind s. Yes, I am not. Bio no loads percentages. Well. OK. But s put ten to if you want to let you know that we do have funds have been around from many years that has shown. Great potential that's shown twenty we did assign that really. Well, rating the industry and. Well, with to started I have a couple funds that you know you gave me for nine into the bond fund and you call bond fund hasn't been meeting your needs and have enough to bad taste in you now. Well, Premium see you talking about because fund that I can make you or for your mom I know that. But I Simply telling you have a quick couple of Monday's with you Mr bee this strategic Income. Right on that right now the venue Iomega I'll have a lot of strategy in the Strategic Income. OK, that I feel with every you out that maybe everybody does have underperforming funds in the portfolio and is there a busy part of your underperforming portfolio I you feel about the middle bond fund we do have a big position and that has been doing you have conventional we are. You sure that been not performance we also you what they say you know. Well, I can always you know I looked into something newspaper we show the one year performances on a is there afford to be dreadful by fund also. Yeah, that has Sweat what. You have a Minus for the one year before that you look have all along are. Right you feel that has performed for you right now were Term I would say average before if there's I would still say to go over average top ten we what you what his you consider that man. Yeah, time you want my name and then I'm sorry a blue chip fund number's one because the middle bond fund over ten year period has returned ans. Over a five year period of five point eight. Are not competitor with Dreyfus fund you. Well, I think it's very compatible considering that besides getting the tax transaction is showing about an average return by bond funds. Hello, can I Manhattan compare with a People's for s in any way out to look that up sir. Bye, going to different screaming get you that information of your call I can also the value that the central bond fund that you that you have a phone number you also has shown over ten years eight point seven and you have to five years eight point seven prospectus sit it to look at before last year for ninety nine Diego. I you know the bottom turn five point seven the central bond fund maybe four point IRA I and of course is what bond is taxable see so maybe the central bond is ninety Venue with the most benefit as you possibly maybe furious you're in a fund that's giving you five point six

Figure 6 - A Lightweight playback client using DHTML. Some terms were modified as before.

8. Measuring Accuracy Using Salient Terms

If you look at the absolute word recognition accuracy of these telephone transcriptions, you would find it to be as low as 20-30% in some cases, which at first seems a depressingly low number. This low accuracy is a result of the informal speech, incomplete sentences, regional accents and careless use of grammar and even pronunciation that frequently occur in telephone calls.

We selected the particularly problematic document described above for further analysis, and carefully transcribed the actual conversation, omitting stalling sounds such as “um” and “uh,” and corrected for or omitted any overlapping dialog. We then ran this document through Textract as part of the collection of the remaining 522 documents, so that the same canonical forms could be developed in term recognition for this document.

After loading the results of Textract into our DB2 database, we compared the terms found in the recognized and manually transcribed documents. The results are shown in Table 2.

	manual	Auto-matic	%
Multiwords found in document	22	12	55
Multiwords less spurious finds	19	12	63
Multiwords where singles found	19	13.5	71

Table 2 –Recognition of Multiwords in a manually transcribed and automatically recognized document.

Of the several hundred terms found in both documents, we tabulated the multiword terms found in both. For the most part, these multiwords represent the high IQ salient concepts that are the highlighted terms in our display, and serve as a thumbnail outline of the conversation. There were 22 such terms in the manually transcribed document, where the opportunities for forming words correctly without intervening noise were greatest. Of those, 12 were found in the speech-recognized document.

However, once you eliminate multiwords caused by Textract “misfires” and the customer’s name, which the recognition engine could not be expected to get, the correct rate rises to 12 out of 19. Finally, if you give half-scores if both words of the multiword are found individually, we get 13.5 out of 19 or 71% accuracy in finding the salient concepts in the document.

Thus, we find, that even when very difficult conversations are subjected to speech recognition, the speech engine is quite capable of finding the preponderance of the salient terms in a document, and text mining systems like Textract are very capable of extracting these concepts and using them to provide note-like summaries of the major concepts in those conversations. In fact, even with extremely problematic recognition, caused by careless speakers and difficult regional accents, the ability of the speech engines to provide valuable document summaries remains very strong.

9. Results and Discussion

In these experiments, we discovered that while it is not yet possible to produce word-accurate transcripts of informal telephone conversations, it is still possible to provide extremely useful information regarding the content of these calls. Rather than regarding the recognized text as a call transcript, it is more useful to consider it a beginning of a set of “notes” on the call such as either party might take down while talking to remind themselves later of the conversation.

It is not yet possible to provide the sales information that was projected before we began the project. However, we did find that by recognizing the salient terms in the conversation and providing a search system for searching the call archive, we can provide supervisors with a way of looking into what information was discussed, and a way of playing back the interesting portions of calls returned from such a search.

Finally, we found that while the word accuracy calls of these transcripts was in many cases fairly low, the salient term accuracy was quite high and made these searchable summaries extremely useful.

Acknowledgments

We'd like to thank Branimir Bogaraev for helpful suggestions which led to the development of the client display, Yael Ravin for modifications to Textract to make name detection more reliable in this environment, Paul Kaye for writing the media player interface, Bob Mack for helpful general suggestions, Salim Roukos for helpful conversations, Harry Kolar for thoughtful analysis of the results, Michael Hehenberger for his work in developing the project, and Roy Byrd, Alan Marwick for their technical support.

Bibliography

Brandow, Ron, Karl Mitze, and Lisa Rau, 1995. Automatic condensation of electronic publications by sentence selection. *Information Processing and Management*, 31, No. 5, 675-685.

Byrd, R.J. and Ravin, Y. Identifying and Extracting Relations in Text. *Proceeedings of NLDB 99*, Klagenfurt, Austria.

Buckley, C., Singhal, A., Mira, M & Salton, G. (1996) "New Retrieval Approaches Using SMART:TREC4. In Harman, D, editor, Proceedings of the TREC 4 Conference, National Institute of Standards and Technology Special Publication.

Chen, Scott Shaobing, M.J.F. Gales, P.S. Gopalakrishnan, R.A. Gopinath, H. Printz, D. Kanevsky, P. Olsen, L. Polymenakos, IBM's LVCSR System for Transcription of Broadcast News Used in the 1997 Hub-4 English Evaluation in Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, 1998

Cooper, J.W. and Prager, John M, *Anti-Serendipity: Finding Useless Documents and Similar Documents*, Proceedings of HICSS-33, Maui, HI, January, 2000.

Cooper, J.W., "Colorful Language," JavaPro **4**(1), 44, 2000.

Cooper, James W. and Byrd, Roy J. "Lexical Navigation: Visually Prompted Query Expansion and Refinement." Proceedings of DIGLIB97, Philadelphia, PA, July, 1997.

Cooper, James W. and Byrd, Roy J., OBIWAN – "A Visual Interface for Prompted Query Refinement," Proceedings of HICSS-31, Kona, Hawaii, 1998.

Dharanipragada, S. and S. Roukos, Experimental Results in Audio Indexing in Proceedings of the DARPA Speech Recognition Workshop, 1997

Fowler, Richard H., Wilson, Bradley A., and Fowler, Wendy A.L. "Information Navigator: An information system using networks for display and retrieval." Report NAG9-551, No.92-1. Department of Computer Science, University of Texas, Pan American, Edinburg, TX.

Johnson, S.E., Jourlin, P., Spark-Jones, K. and Woodland, P.C. *Spoken Document Retrieval for TREC-8 at Cambridge University*, Proceedings of TREC-8, 1999.

McGill University FTP Archive. See ftp.TSP.ECE.McGill.CA/pub/Afsp

Neff, Mary S. and Cooper, James W. 1999a. Document Summarization for Active Markup, *in Proceedings of the 32nd Hawaii International Conference on System Sciences*, Wailea, HI, January, 1999.

Pekowsky, Larne. *JavaServer Pages*, Addison-Wesley, Boston, MA, 2000.

Prager, John M., Linguini: Recognition of Language in Digital Documents, *in Proceedings of the 32nd Hawaii International Conference on System Sciences*, Wailea, HI, January, 1999.

Schatz, Bruce R, Johnson, Eric H, Cochrane, Pauline A and Chen, Hsinchun, "Interactive Term Suggestion for Users of Digital Libraries." *ACM Digital Library Conference*, *1996.*

Viswanathan, M, H.S.M. Beigi, S. Dharanipragada, and A. Tritschler, "Retrieval from Spoken Documents Using Content And Speaker Information," Proceedings, International Conference on Document Analysis and Retrieval (ICDAR99), Bangalore, India, 1999, pp. 567--572.

Woodland, P.C., Johnson, S.E., Jourlin, P. and Sparck-Jones, K. *Effects of Out of Vocabulary Words in Spoken Document Retrieval*. Proceedings of SIGIR 2000, Athens, Greece, July, 2000.

Xu, Jinxi and Croft, W. Bruce. "Query Expansion Using Local and Global Document Analysis," *Proceedings of the 19th Annual ACM-SIGIR Conference*, 1996, pp. 4-11