MFCC Based Speech Retrieval

Jyoti Srivastava, Tanveer J. Siddiqui, U. S. Tiwary, Ashish Kumar Srivastava

Abstract: This paper presents an approach for speech retrieval. The feature being used in this approach is MFCC. This approach does not use any phoneme recognizer or Speech to text tool hence it can be used for other languages as well leads to the problem of speech retrieval (SR). This method retrieves ranked audio files containing spoken text in response to a given speech query. In this paper indexing methods are described which represent the contents of the spoken documents. The indexing methods, which are based on the output of phoneme recognizer, take account of speech recognition errors. While in this paper, speech documents are directly compared with the speech query based on MFCC. Thus, reduced the overhead of conversion from speech to text.

Index Terms: MFCC, Phoneme recognizer, Speech Retrieval, Speech comparison.

I. INTRODUCTION

There is vast amount of data available in audio form like reports, broadcast news, interviews, documentation, discussions, radio plays, recorded lectures etc. It is important to store and retrieve this information in response to the user's query.

A lot of research work has been done on speech retrieval system. Most of these works used various types of transcribed units for indexing. In these approaches the performance of system depends on the performance of the speech recognizer tool.

A speech retrieval system accepts vague speech queries and it performs best-match searches to find audio files that are likely to be relevant to the queries.

The speech queries and spoken documents must be converted into content features such as keywords, phone strings, and texts using speech recognition techniques. An alternative approach is to use the MFCC coefficients for indexing. These MFCC coefficients are used to measure the similarity between the speech queries and the spoken documents. Selecting appropriate content features to represent the spoken documents and speech queries is thus very important.

The main dissimilarity to text retrieval is to cope up with the following problems:

- Word boundaries are difficult to detect. Every speech interval system may represent occurrence of an indexing feature.
- Recognition errors (non-detections) affect the retrieval effectiveness.
- The size of the indexing vocabulary is limited by the number of different words that can be recognized by a speech recognition system.

These problems raise the questions of what indexing features should be used to describe the content of the speech documents. In text retrieval, all words which are not too common (i.e. all but stop words) are selected as indexing features. Such a selection of indexing features is not feasible in speech retrieval because of the limitations of current speech recognition systems.

Spoken information is also of growing interest in research and education where for example talks held at conferences or lectures presented in universities are recorded in order to make them available for other research sites and students respectively. These developments are further supported by improved networking environments such as the World Wide Web (www) where vast amounts of data are made publicly available.

Finally spoken information in digitized format starts to play a major role in private and business communications in the form of voice or video mail messages. Some of the Applications of speech retrieval are:

- Retrieval of music or sounds from large archives.
- Retrieval of videos by their sound track.
- Classification of music and sounds by similarity.
- Monitoring phone conversations.
- Recorded lecture retrieval.
- News retrieval.
- Useful in question answering system.

Figure II.1 shows the architecture of the speech retrieval system in which there is a database which contains the entire spoken documents. A user can enter the query according to their information need into the system in the form of speech through the microphone or the user can give already recorded query and then the system will retrieve the audio files relevant to the query. Now the user can select and play any of the retrieved relevant audio file.

II. LITERATURE SURVEY

Speech retrieval system accepts user's query in the form of speech and returns a list of relevant documents. It is a content-based retrieval. Indexing generates content- based description of the documents. Units used for indexing are called indexing features. Previous approaches for the speech retrieval used speech recognition in the first step. After recognizing the speech, they apply text retrieval techniques.

Revised Manuscript Received on July 05, 2019

Dr. Jyoti Srivastava, Department of Computer Science and Engineering, Madanapalle Institute of Technology & Science, Madanapalle, Chittoor District, Andhra Pradesh, India.

Dr. Tanveer J. Siddiqui, J. k. Institute of Applied Department of Electronics & Communication, University of Allahabad, Allahabad, India.

Prof. U. S. Tiwary, Indian Institute of Information Technology Allahabad, India.

Dr. Ashish Kumar Srivastava, Department of Computer Science and Engineering, Madanapalle Institute of Technology & Science, Madanapalle, Chittoor District, Andhra Pradesh, India.

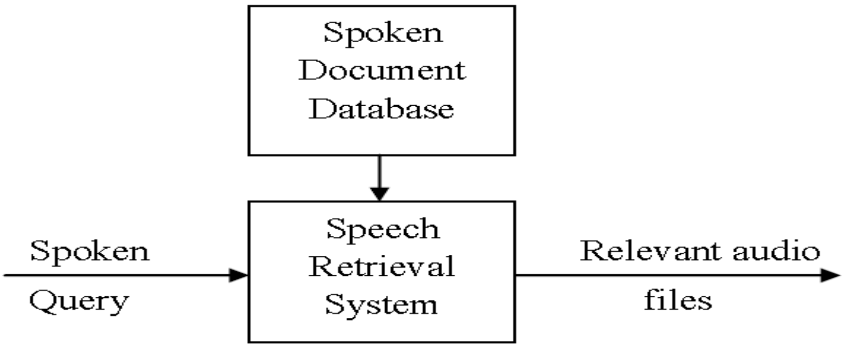

Figure II.1: System Architecture of Speech Retrieval

Speech retrieval thus becomes a two-step process:

- Speech is converted to text.
- Text retrieval technique is applied to retrieve the relevant audio files.

An important prerequisite for Speech Retrieval (SR) is a speech recognition system which usually operates on either the word or the phoneme level. The approach described in the previous papers requires a phoneme recognizer or speech to text converter, which initially generates phoneme sequences or text from the spoken documents. Partly this requirement originally came from our objective to study SR experimentally on documents spoken in English, and from the fact that a suitable speech recognizer for English was not available at the beginning of this work.

From the information retrieval perspective, the main argument in favor of phoneme recognition is that it allows queries from an unrestricted vocabulary. This is allowed because the recognition output (i.e. the phoneme sequences) is not bound to any vocabulary, as opposed to word-recognition-based SDR, where the recognition vocabulary defines and currently restricts the query vocabulary.

A. Approaches for Speech Retrieval

Speech retrieval methods are based on the features extracted from the spoken documents. There can be a number of choices for features.

- **Based on Word Recognition:** This approach converts both the speech documents and query into text so that words can be used as feature to index the documents and then apply any text retrieval techniques to retrieve the relevant documents. A coupling of word recognition and text retrieval was first presented in the video mail retrieval project at Cambridge University [1]. For this approach STT (Speech to Text) tool is required. This approach has two drawbacks [1]. First is that it faces the problem of limited size of recognition vocabulary which directly restrict the query vocabulary. Second is that it needs a huge amounts of training data that contain several occurrences of recognition-vocabulary words.
- **Based on Sub-Word Recognition:** This approach work on the sub-word units recognized from the speech. In [2] a VCV feature is used as recognizable sub-word unit. A VCV feature is a three concatenated sequence of vowel, consonant and vowel, respectively. The recognition system generates a sequence of VCV feature for each spoken document, which is used to create the document description for the retrieval. Bo-Ren Bai and Berlin [3] and Martha, Stefan [4] used the syllable as the indexing feature to index the spoken document for spoken document retrieval. Sub-word-based approaches suffer from two major drawbacks. First is that recognition quality degrades for shorter units because they contain fewer pieces of phonetic evidence. Second is that the features are selected from text without taking their acoustic property into account.
- **Based on Phoneme recognition:** This approach converts both the speech documents and query into phonemes and then applies probabilistic string matching. A phoneme is the basic speech unit. For this approach we need a phoneme recognizer tool. A phoneme recognizer transcribes digitized speech into a sequence of phonemes. This sequence is then used to index the documents. We can use N-gram approach to index the phoneme strings extracted from the speech documents. In this approach query vocabulary is unrestricted so it leads to the open vocabulary retrieval. The effectiveness of the system depends on the phone error rate of the phoneme recognizer.
- **Based on speech feature (MFCC):** This is the approach used in the paper. In this approach we extract the important features from the speech like MFCC (Mel frequency cepstrum coefficient) and then apply coefficient matching. Silent portion is removed from the speech before extracting the feature, because it is not the useful information for retrieval but increase the computation cost. This approach will work well if the coefficient of different audio files having the same utterances is the same. This approach is also independent of the recognition vocabulary. This is the approach followed by this paper.

B. Problems with the Speech Documents

The main problem while applying speech recognition for spoken document retrieval is the quality or accuracy of the recognition output. Accurate recognition results lead to higher quality document descriptions and thus better retrieval effectiveness. However, ASR is a difficult task and accordingly its output often contains a considerable number of recognition errors. The recognition output quality is mainly affected by the following factors [5]:

- *Speech variability:* The temporal and acoustic properties of the same utterance may vary considerably, even if the same text is spoken.

- *Speech type:* Continuous speech is more difficult to recognize than if words were spoken in an isolated manner because no explicit word boundaries are present.
- *Number of distinct units to recognize:* Defining larger sets of recognizable units increases complexity, and thus the risk that units may be confused during recognition.
- *Amount and the quality of training data:* These factors are necessary to train both acoustic models to recognize individual units (e.g. words), and language models to define possible combinations of recognizable units in continuous speech.
- *Number and gender of different speakers:* Both speaking speed and pronunciation are individual to a speaker. Speaker independent recognition is far more difficult than speaker dependent recognition.
- *Recording environment:* Background noise usually affects the analysis of the actual speech signal. Note that background noise may arise in the environment of the speaker or in the communication channel (e.g. in a telephone line).

It may generally be said that recognition errors may considerably affect retrieval effectiveness.

C. Performance Comparison of Text Retrieval and Speech Retrieval

Spoken document retrieval system for different languages faces the problem because it handles with the speech rather than the text.

- Performance of German Spoken Document Retrieval system is degraded due to poor performance of phoneme recognizer. Compared with the text retrieval, the effectiveness of the best PSM (Probabilistic string matching) method was found to decrease by 43% in terms of average precision. However, this result is based on a very poor phoneme recognizer with a 59% phoneme error rate [5].

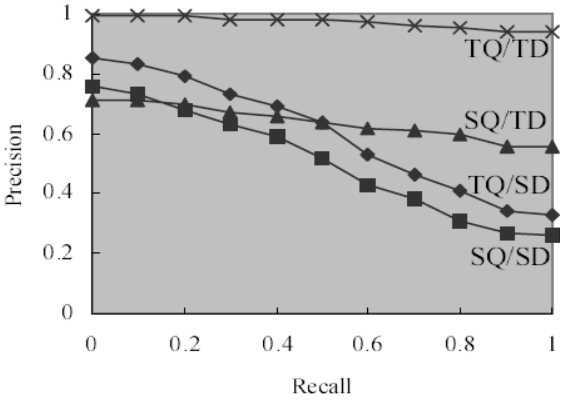

Figure II.2: Performance of the Spoken/Text Document Retrieval for Speech/Text Queries

- Chinese Spoken Document Retrieval system uses the syllables as the feature to index the spoken documents. It retrieves the Text/Spoken documents using the Text/ Spoken queries. The large difference in performance of both can be easily detected when we use documents and query in speech form rather than the text [3]. So, the performance of the text retrieval is always high than the performance of the spoken document retrieval.

III. PROPOSED APPROACH

The approach that is used in this paper for speech retrieval is based on speech features. The speech feature used to index the speech is MFCC coefficients because in literature review it is analyzed that MFCC coefficients is the most popular feature used for speech recognition. Before extracting the feature, the silence portion is removed from the speech file, because it is not so much important for retrieval. And working on speech file including silence portion just increase the computation cost and time. After removing the silence portion, file size is reduced up to 50% and more.

The work of this paper can be described in the following steps:

A. Collect the Data Set

Speech data is collected from CMU site of the size 210 MB. It is distributed into 1300 different speech files. Speech files are in .wav format. Each different file size is varying form 65 KB to 500 KB. I have recorded my own data set too, which have the information regarding news. The news taken from different newspapers and recorded it. I recorded 100 documents and 50 queries.

B. Remove the Silence Part of the Speech File

Silence part of the file is not so much important in the case of speech retrieval. Moreover, it increases the file size and has no useful information. Processing of the speech file with silence part increases the computation cost. So, it is better to remove it. It is analyzed that which parts of the speech file contains the silence information, by seeing the waveform of different sentences. Silence part is removed based on the amplitude information of the speech waveform, the amplitude at that time in the waveform is in between -0.05 to +0.05, but it is not true for all speech file. The range of the silence portion for different file may vary. So, it is better to take the range for silence is as in between maximum value of the amplitude/4 to the minimum value of the amplitude/4. It works well for almost each file.

The algorithm used to remove the silence of the speech file is as follows:

```
1. Read the speech waveform and put it into an array (say W) .
2. M = min (W)/4.
3. N = max (W)/4.
4. Remove W lies in between M and N remove it.
```

5. The resulting W is without silence portion. Figure III.2 shows the effect of removing silence portion from Figure III.1.

DOI: 10.35940/ijitee.I7550.078919

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

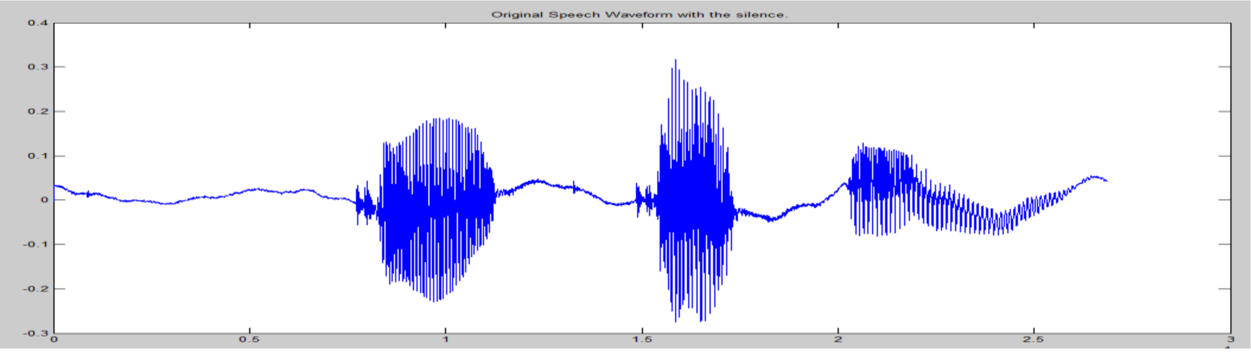

Figure III.1: Original waveform before removal of silence portion

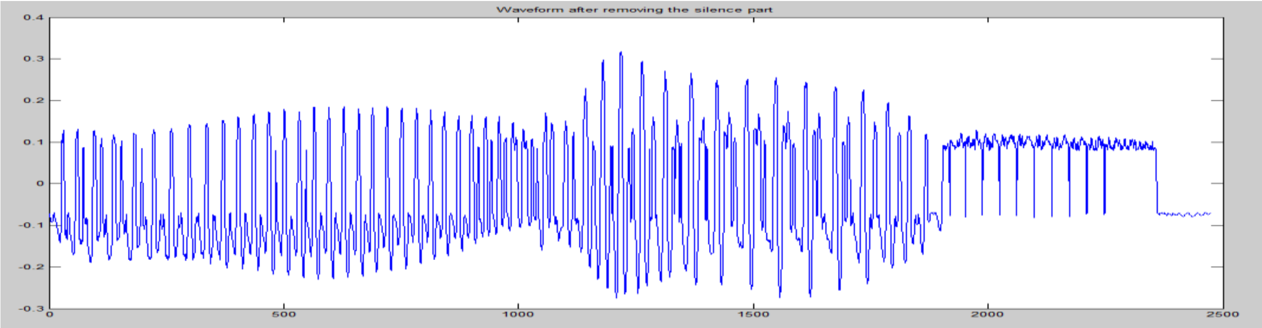

Figure III.2: waveform after removal of silence portion.

C. Calculate MFCC Coefficients for Each Frame

In speech processing, the Mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency. Cepstrum is also a good feature of speech but in this paper, Mel-frequency Cepstrum is used. The reason is that MFC is better than cepstrum in the sense that in the MFC the frequency bands are equally spaced on the Mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstrum. Thus, MFC can allow for better representation of sound. It is the reason that MFC used mainly in speech recognition.

D. Process to calculate MFCC Features

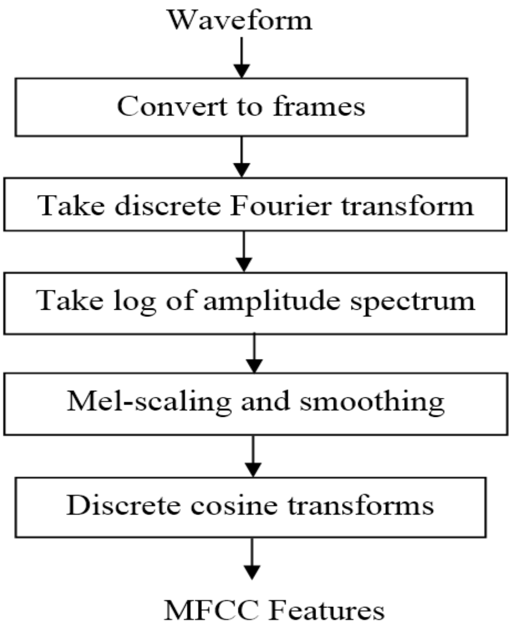
Waveform
Convert to frames
Take discrete Fourier transform
Take log of amplitude spectrum
Mel-scaling and smoothing
Discrete cosine transforms
MFCC Features

Figure III.3: Flow chart to calculate MFCC features

IV. SIMULATION RESULT AND ANALYSIS

This section evaluates the effectiveness of proposed approach for speech retrieval. This paper uses a collection of documents recorded in English.

Table IV.1: Properties of the test collection

Total collection duration	8 min
Duration of documents	5s – 10s
Number of documents	100
Number of spoken words per documents	8 - 20
Number of queries	25
Number of words per query	≈ 3
Number of relevant documents per query	≈ 4

A dataset is constructed to test this speech retrieval system. So, this dataset has the news taken from different website. This system is tested on 100 documents and 25 queries. In which each query is related to approximately 4 documents. The collection contains the 100 news headlines recorded. Properties of this collection are summarized Table IV.1.

A. System Performance for The Best Case, Average Case and Worst Case

The best case, average case and worst-case result of my speech retrieval system in the recall/precision graph are plotted in the Figure IV.1. In the best-case average precision for the system is 0.3177 and for the average case the average precision for the system is 0.2310 and for the worst case the average precision for the system is 0.0255.

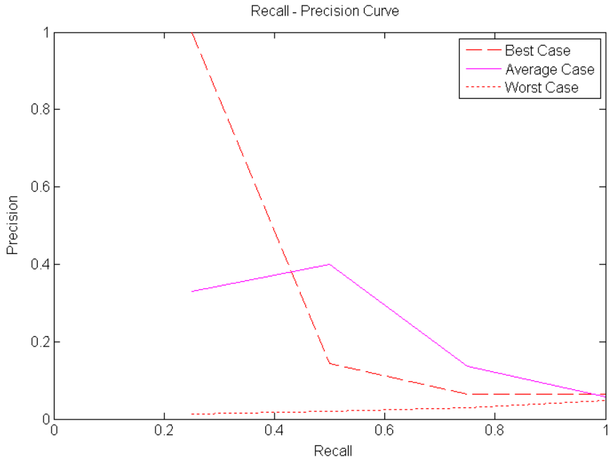

Figure IV.1: Recall/Precision curve for the best case, average case and worst case

B. Interpolated Precision/Recall graph

Precision-recall curves have a distinctive jagged shape. If the next document retrieved is not relevant then recall is the same, but precision has dropped. If it is relevant, then both precision and recall will increase, and the curve jags to the right. The standard way to remove these jiggles is through interpolated precision. The interpolated precision at a certain recall level r is defined as the highest precision found for any recall level greater than or equal to r.

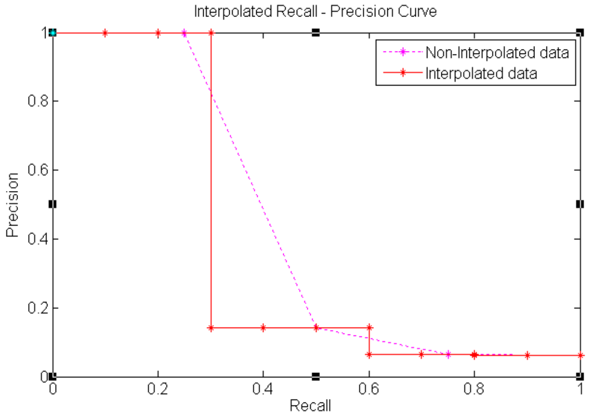

Figure IV.2: Interpolated Precision/Recall graph

In Figure IV.2 non-interpolated precision is plotted with the dotted line and interpolated precision is plotted with the solid line where '*' represent the precision point.

C. Precision at 11 Standard Recall levels

The precision averages at 11 standard recall levels are used to compare the performance of different systems and as the input for plotting the recall-precision graph. Each recall-precision is computed by summing the interpolated precisions at the specified recall cut-off value and then dividing by the number of queries used to test the system. For the precision-recall curve in Figure IV.2, these 11 values are shown in Table IV.2.

For all 11-standard point of recall level, we then calculate the arithmetic mean of the interpolated precision at that recall level for each query in the test collection. A composite precision-recall curve showing 11 points of precision for these 11 standard recall levels can then be graphed.

Table IV.2: Calculation of 11-point Interpolated Average Precision

Recall	Interpolated precision
0.0	0.1463
0.1	0.1463
0.2	0.1463
0.3	0.0908
0.4	0.0908
0.5	0.0908
0.6	0.0717
0.7	0.0717
0.8	0.0559
0.9	0.0559
1.0	0.0559

Figure IV.3 shows such an Averaged 11-point precision/recall graph over 25 queries for our speech retrieval system. Non-interpolated average precision over all relevant documents is 0.0559.

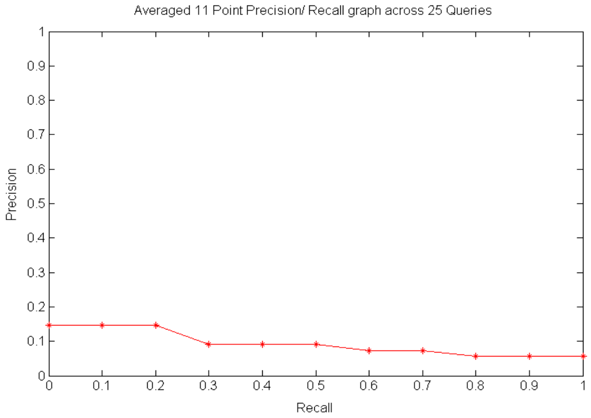

Figure IV.3: Averaged 11 Point Precision/Recall graph across 25 queries

V. CONCLUSION AND FUTURE WORK

Proposed approach is not using any language model or anything specific to a language, it is just taking coefficient form one speech and match it with coefficient of another speech and it will match if there will be some similar kind of phonemes in both speech file. So, we can use it for any language but not for cross language. And we have no need to store a large vocabulary too. So, this speech retrieval approach has two main advantages due to this approach:

- Independent of the language.
- Independent of the size of vocabulary.

In comparison to text retrieval system the performance of the speech retrieval system is not much good as studied into the literature survey part also. It can be improved by doing some modification in it. As the system used the amplitude of the speech signal as the feature to remove the silence part of the speech file, one can try the energy thresholding method to remove the silence portion. It needs a lot of time for experimentation.

In future, we can try to work and apply these techniques to improve the performance of the speech retrieval system.

REFERENCES

1. Jones, G., Foote, J., Jones, K. S., & Young, S., "Video Mail Retrieval using Voice: An Overview of the Stage 2 System." Presented at MIRO Workshop, Glasgow, September 1995.
2. Glavitsch, U., & Schauble, P., "A System for Retrieving Speech Documents", ACM SIGIR conference on R&D in Information retrieval, pp. 168-176, 1992.
3. Bo-Ren Bai, Berlin Chen and Hsin-Min Wang "Syllable-Based Chinese Text/Spoken Document Retrieval Using Text/Speech Queries", WSPC/115-IJPRAI, August 2000.
4. Christian Schrumpf, Martha Larson, and Stefan Eickeler, "Syllable-based Language Models in Speech Recognition for English Spoken Document Retrieval" presented in International Workshop on Audio-Visual Content and Information Visualization in Digital Libraries, May 2005.
5. Martin Wechsler, P. Schauble, C. J. van Rijsbergen, "Spoken Document Retrieval Based on Phoneme Recognition", DISS. ETH No. 12879, 1998.
6. Huixiang Gu, Jianming Li, Ben Walter, Eric Chang, "Spoken Query for Web Search and Navigation", Poster Proc. 10th Int. World-Wide Web Conf, 2001.