188.412 VU 3.0 Information Retrieval 14.06.2007

1. Music-Retrieval: Beschreiben sie den Prozess zur Berechnung der Statistical Spectrum Descriptors sowie den Unterschied zu Rhythm Histograms betreffend der Art der Information die dargestellt wird. (5 Punkte)

SSD-Berechnung:

Zuerst wird wie bei RP (Rhythm Patterns) ein Sonogramm erstellt (mit Hilfe der „Short Time Fourier Transform“):
P1: Konvertiere Datei in unkomprimiertes Digital Audio
P2: Konvertiere zu Mono
P3: Nimm 6sec Ausschnitte
S1: Erstelle Spektogramm (mit FFT)
S2: wende Bark-Skala an (24 Critical-Bands)
S3: Anwenden einer Spreizfunktion wegen Maskierungseffekten
S4: Transformiere Spektrumswerte in dB
S5: Errechne Lautstärken-Levels in Phon
S6: Errechne spezifische Lautstärkenwahrnehmung in Sone
Dann werden für jedes „Critical Band“ Statistische Maße (Durchschnitt, Median, Varianz, Schiefe, Wölbung, Min, Max) errechnet.
Abschließend wird der Median über die 6sec Abschnitte bestimmt.

RH:

Aus dem Sonogramm wird durch eine Fourier Transformation eine Zeininvariante Repräsentation sich wiederholender Abschnitte erzeugt.
Im Gegensatz zu SSD und RP wird die Information nicht per „Critical Band“ gespeichert sondern per Modulationsfrequenz. Die Höhe aller „Critical Bands“ für eine Frequenz werden pro Segment summiert und bilden ein Histogramm der „rhythmischen Energie“. Für ein ganzes Musikstück wird der Median der Histogrammwerte in den einzelnen Segmenten gebildet.

2. Text-Indexing: Beschreiben sie unterschiedliche Methoden zur Feature Space Reduction (Pruning) bei hochdimensionalen Featurevektoren für überwachte und unüberwachte Textanalyse-Aufgaben. (10 Punkte)

Pre-Processing:

Case Folding : Groß ODER Kleinschreibung, nicht beides.
Collection Cleansing: Unbrauchbare Information entfernen (Formatierung, …)
Stemming: Reduziert die Wörter auf ihren Wortstamm (linguistic in- / correct)
Stop-Word Removal: Entfernt sehr häufige Wörter (wenig Informationsgehalt)
Remove Very Rare Terms: viele Worte sind nur selten → gute Kompression

Indexing:

N-Gram: Darstellung in Abschnitten bestimmter Länge (+Robust, -Effizienz)
Semantic Concepts: Spezielle Konzepte für z.B. Zeitangaben, Orte, Personen, …
Phrases Co-occurences:
Word Stems: Verwendet die Wortstämme statt den Termen

3. ATC: Beschreiben sie die Qualitätsmaße Precision und Recall, sowie den Unterschied zwischen den Micro/Macro Formen derselben zur Evaluierung eines Classifiers, wie sie ermittelt werden, und was sie aussagen. (10 Punkte)

Micro Averaged:

$$\mathrm{Precision} = \frac{\sum TP}{\sum (TP + FP)}$$

$$\mathrm{Recall} = \frac{\sum TP}{\sum (TP + FN)}$$

Macro Averaged (Zuerst per Kategorie, dann Mitteln):

$$\mathrm{Precision} = \frac{\sum \mathrm{Precision}(i)}{Anz.Kategorien}$$

$$\mathrm{Recall} = \frac{\sum \mathrm{Recall}(i)}{Anz.Kategorien}$$

Precision gibt die Wahrscheinlichkeit an mit der ein gefundenes Objekt relevant ist. Ist das Verhältnis der richtig Positiven zu allen gefundenen Objekten. (positiver Vorhersagewert, Relevanz).

Recall gibt die Wahrscheinlichkeit an mit der ein relevantes Objekt gefunden wird. Ist das Verhältnis der richtig Positiven zu allen eigentlich relevanten Objekten. (Sensitivität, Empfindlichkeit)

Es ist kaum möglich beide Maße gleichzeitig zu optimieren, denn eine hohe „Precision" erhöht gleichzeitig die Wahrscheinlichkeit seltene Fälle nicht abzudecken. Und bei der Optimierung von „Recall" wird es dazu führen auch nicht relevante Objekte einzubeziehen.

Unterschied ??

4. Information-Extraction: Was macht das Lexicon & Morphology Modul? Wann ist eher die Verwendung eines Lexicons, wann eines Morphology Moduls angebracht? (10 Punkte)

Im „Lexicon & Morphology Modul" werden die Tokens (vom „Tokenization Modul") auf ihre Wortart (Verb, Nomen, Adjektiv, …) überprüft. Dies kann durch einfaches Nachsehen in einem Lexikon oder durch ein morphologisches Modul („Part-of-speech Tagger") durchgeführt werden. Das morphologische Modul versucht die Bedeutung und Funktion eines Wortes durch seine Morpheme (kleinster Wortteil) zu ergründen.Die Wichtigste Aufgabe des L&M Moduls ist die richtige Zuordnung von Eigennamen.

Ein Lexikon kann eingesetzt werden wenn alle Varianten eines Wortes leicht aufgelistet werden können (z.B. im Englischen). Wenn hingegen die Komplexität der Sprache diese Auflistung zu einer schwierigen oder gar unlösbaren (z.B. im Deutschen, wo durch Komposition mehrerer Wörter neue Wörter entstehen) Problematik werden lässt muss auf die Morphologie zurückgegriffen werden.

5. Text summarization: Was versteht man unter abstraktiver bzw. extraktiver Summarization? (10 Punkte)

Extraktive Extraktion: Relevante Sätze oder Paragraphen werden gesucht. Die gefundenen Stellen werden gereiht und in eine Zusammenfassung kopiert. Morphologische Analyse ist nicht notwendig.

Abstraktive Extraktion: Verwendet entweder Vorwissen über die Struktur der verlangten Information („Domain knowlege") und füllt die Information in „Scripts" oder „Templates", welche dann durch NLG-Systeme in zusammenfassende Sätze transformiert wird.
Oder es wird eine Semantische Repräsentation des Dokuments „gebaut" welche dann wieder mit NLG-Systemen in eine leserliche Form gebracht werden kann.

6. QA: Nennen sie die gebräuchlichsten Komponenten eines Question Answering Systems. (5 Punkte)

Analyze Question
Gather information
Distill answers
Sanity Check
Present answers

188.412 VU 3.0 Information Retrieval 31.10.2007

1. Music-Retrieval: Beschreiben sie den Prozess zur Berechnung der Statistical Spectrum Descriptors sowie den Unterschied zu Rhythm Histograms betreffend der Art der Information die dargestellt wird.

SSD-Berechnung:

Zuerst wird wie bei RP (Rhythm Patterns) ein Sonogramm erstellt (mit Hilfe der „Short Time Fourier Transform"):
P1: Konvertiere Datei in unkomprimiertes Digital Audio
P2: Konvertiere zu Mono
P3: Nimm 6sec Ausschnitte
S1: Erstelle Spektogramm (mit FFT)
S2: wende Bark-Skala an (24 Critical-Bands)
S3: Anwenden einer Spreizfunktion wegen Maskierungseffekten
S4: Transformiere Spektrumswerte in dB
S5: Errechne Lautstärken-Levels in Phon
S6: Errechne spezifische Lautstärkenwahrnehmung in Sone
Dann werden für jedes „Critical Band" Statistische Maße (Durchschnitt, Median, Varianz, Schiefe, Wölbung, Min, Max) errechnet.
Abschließend wird der Median über die 6sec Abschnitte bestimmt.

RH:

Aus dem Sonogramm wird durch eine Fourier Transformation eine Zeininvariante Repräsentation sich wiederholender Abschnitte erzeugt.
Im Gegensatz zu SSD und RP wird die Information nicht per „Critical Band" gespeichert sondern per Modulationsfrequenz. Die Höhe aller „Critical Bands" für eine Frequenz werden pro Segment summiert und bilden ein Histogramm der „rhythmischen Energie". Für ein ganzes Musikstück wird der Median der Histogrammwerte in den einzelnen Segmenten gebildet.

2. Text-Indexing: Beschreiben Sie unterschiedliche Methoden/Arten von Feature Spaces die bei der Indizierung von Textdokumenten zum Einsatz kommen, wie sie aufgebaut werden, sowie ihre Eignung bzw. Vor/Nachteile. Beschreiben sie weiters die tfxidf Methode zur Termgewichtung, wie diese Familie prinzipiell berechnet wird, und welche Annahmen dahinter stecken.

Feature Spaces:

Bag of Words „BOW"(Liste aller Terme): Verbreitetste, Unabhängigkeits-vermutung, Hochdimensional,
N-grams (Sequenz von n Buchstaben): Kein Stemming notwendig, Features nicht intepretierbar
Word/-stemms (Wörterliste): Verbreitetste,
Co-Occurrences (Wiederholungen):
Semantic Concepts (dates, locations, persons, ..): braucht NatLangProcessing, sehr Domänen spezifisch

tf (Text-Freq): Anzahl der Vorkommen eines Wortes / Anz Wörter.
Je öfter ein Term in einem Dokument vorkommt, desto wichtiger ist er. Bedeutung des Terms für das Dokument.

idf (Invers-Document-Freq): ln(Anz.Dok mit diesem Wort / Anz. Dok in Kollektion)
Je mehr Dokumente in einer Kollektion diesen Term beinhalten, desto unwichtiger ist er. Bedeutung des Terms für Kollektion.

tfxidf = tf*1/df (= tf/df) ➔ statt idf auch oft ln(N(Dok)/df)

3. ATC: Beschreiben sie die Qualitätsmaße Precision und Recall, sowie den Unterschied zwischen den Micro/Macro Formen derselben zur Evaluierung eines Classifiers, wie sie ermittelt werden, und was sie aussagen.

Micro Averaged:

$$\mathrm{Precision} = \frac{\sum TP}{\sum (TP + FP)}$$

$$\mathrm{Recall} = \frac{\sum TP}{\sum (TP + FN)}$$

Macro Averaged (Zuerst per Kategorie, dann Mitteln):

$$\mathrm{Precision} = \frac{\sum \mathrm{Precision}(i)}{Anz.Kategorien}$$

$$\mathrm{Recall} = \frac{\sum \mathrm{Recall}(i)}{Anz.Kategorien}$$

Precision gibt die Wahrscheinlichkeit an mit der ein gefundenes Objekt relevant ist. Ist das Verhältnis der richtig Positiven zu allen gefundenen Objekten. (positiver Vorhersagewert, Relevanz).

Recall gibt die Wahrscheinlichkeit an mit der ein relevantes Objekt gefunden wird. Ist das Verhältnis der richtig Positiven zu allen eigentlich relevanten Objekten. (Sensitivität, Empfindlichkeit)

Es ist kaum möglich beide Maße gleichzeitig zu optimieren, denn eine hohe „Precision“ erhöht gleichzeitig die Wahrscheinlichkeit seltene Fälle nicht abzudecken. Und bei der Optimierung von „Recall“ wird es dazu führen auch nicht relevante Objekte einzubeziehen.

Unterschied ??

4. Um Extraction Patterns zu definieren gibt es den atomaren und den Molekularen Ansatz. Beschreiben sie die beiden Ansätze. Welcher ist gebräuchlicher? Warum?

Atomarer Ansatz:

Am Anfang sind die Regeln nicht sehr spezifisch und versuchen möglichst alle, eventuell relevante, Information zu extrahieren, dies wird auch „intelligent guessing“ genannt. Dieser Ansatz hat am Beginn einen hohen „Recall“ und niedrige „Precision“. Durch Filtertechniken und Heuristiken kann dann versucht werden die „Precision“ zu steigern.

Molekularer Ansatz:

Dies ist die häufigste Methode. Zuerst werden wenige, sehr verlässliche Regeln verwendet um die wichtigsten Patterns der Domäne zu extrahieren. Diese Regeln werden im Verlauf generalisiert um auch weniger häufige Fälle abzudecken. Dieser Ansatz hat also am Anfang eine hohe „Precision“ und niedrigen „Recall“.

5. Was versteht man unter Coreferencing? Welche Arten von Coreferencing gibt es? Geben sie Beispiele an.

Da in einem Text ein Objekt oft auf unterschiedliche Arten referenziert wird versucht Coreferencing Gruppen von Tokens zu finden die einander referenzieren und damit die selbe Bedeutung haben. Es gibt 3 Arten von CoReferencing:

Name Alias CoReferencing:

Namen und deren Abwandlungen und Abkürzungen sollen erkannt werden.
z.B.: „Mr. Bush“ ist eine Referenz von „George W. Bush“.

Pronoun Antecedent CoReference:

„he, she, it, …“ sind Referenzen auf bestimmte Dinge (z.B. Personen) und müssen zugeordnet werden.

Definite Description CoReference:

Macht nur bei sehr spezifischen Domänen Sinn. Wird verwendet um sehr spezifische Dinge einzuordnen.
z.B.: „Die Tablette“ ist Referenz von „Aspirin“.

6. Information-Extraction: Was macht das Lexicon & Morphology Modul? Wann ist eher die Verwendung eines Lexicons, wann eines Morphology Moduls angebracht?

Im „Lexicon & Morphology Modul“ werden die Tokens (vom „Tokenization Modul“) auf ihre Wortart (Verb, Nomen, Adjektiv, …) überprüft. Dies kann durch einfaches Nachsehen in einem Lexikon oder durch ein morphologisches Modul („Part-of-speech Tagger“) durchgeführt werden. Das morphologische Modul versucht die Bedeutung und Funktion eines Wortes durch seine Morpheme (kleinster Wortteil) zu ergründen.
Die Wichtigste Aufgabe des L&M Moduls ist die richtige Zuordnung von Eigennamen.

Ein Lexikon kann eingesetzt werden wenn alle Varianten eines Wortes leicht aufgelistet werden können (z.B. im Englischen). Wenn hingegen die Komplexität der Sprache diese Auflistung zu einer schwierigen oder gar unlösbaren (z.B. im Deutschen, wo durch Komposition mehrerer Wörter neue Wörter entstehen) Problematik werden lässt muss auf die Morphologie zurückgegriffen werden.

188.412 VU 3.0 Information Retrieval (Rauber) 3.April 2008

1. Music-Retrieval: Beschreiben sie den Prozess zur Berechnung der Statistical Spectrum Descriptors sowie den Unterschied zu Rhythm Histograms betreffend der Art der Information die dargestellt wird.

SSD-Berechnung:

Zuerst wird wie bei RP (Rhythm Patterns) ein Sonogramm erstellt (mit Hilfe der „Short Time Fourier Transform"):
P1: Konvertiere Datei in unkomprimiertes Digital Audio
P2: Konvertiere zu Mono
P3: Nimm 6sec Ausschnitte
S1: Erstelle Spektogramm (mit FFT)
S2: wende Bark-Skala an (24 Critical-Bands)
S3: Anwenden einer Spreizfunktion wegen Maskierungseffekten
S4: Transformiere Spektrumswerte in dB
S5: Errechne Lautstärken-Levels in Phon
S6: Errechne spezifische Lautstärkenwahrnehmung in Sone
Dann werden für jedes „Critical Band" Statistische Maße (Durchschnitt, Median, Varianz, Schiefe, Wölbung, Min, Max) errechnet.
Abschließend wird der Median über die 6sec Abschnitte bestimmt.

RH:

Aus dem Sonogramm wird durch eine Fourier Transformation eine Zeininvariante Repräsentation sich wiederholender Abschnitte erzeugt.
Im Gegensatz zu SSD und RP wird die Information nicht per „Critical Band" gespeichert sondern per Modulationsfrequenz. Die Höhe aller „Critical Bands" für eine Frequenz werden pro Segment summiert und bilden ein Histogramm der „rhythmischen Energie". Für ein ganzes Musikstück wird der Median der Histogrammwerte in den einzelnen Segmenten gebildet.

2. Text-Indexing: Beschreiben sie unterschiedliche Methoden zur Feature Space Reduction (Pruning) bei hochdimensionalen Featurevektoren für überwachte und unüberwachte Textanalyse-Aufgaben.

Pre-Processing:

Case Folding : Groß ODER Kleinschreibung, nicht beides.
Collection Cleansing: Unbrauchbare Information entfernen (Formatierung, …)
Stemming: Reduziert die Wörter auf ihren Wortstamm (linguistic in- / correct)
Stop-Word Removal: Entfernt sehr häufige Wörter (wenig Informationsgehalt)
Remove Very Rare Terms: viele Worte sind nur selten → gute Kompression

Indexing:

N-Gram: Darstellung in Abschnitten bestimmter Länge (+Robust, -Effizienz)
Semantic Concepts: Spezielle Konzepte für z.B. Zeitangaben, Orte, Personen, …
Phrases Co-occurences:
Word Stems: Verwendet die Wortstämme statt den Termen

3. ATC: Beschreiben sie die Qualitätsmaße Precision und Recall, sowie den Unterschied zwischen den Micro/Macro Formen derselben zur Evaluierung eines Classifiers, wie sie ermittelt werden, und was sie aussagen, und wann welche Form eher zur Evaluierung angebracht ist.

Micro Averaged:

$$\mathrm{Pr}ecision = \frac{\sum TP}{\sum (TP + FP)}$$

$$\mathrm{Re}call = \frac{\sum TP}{\sum (TP + FN)}$$

Macro Averaged (Zuerst per Kategorie, dann Mitteln):

$$\mathrm{Pr}ecision = \frac{\sum \mathrm{Pr}ecision(i)}{Anz.Kategorien}$$

$$\mathrm{Re}call = \frac{\sum \mathrm{Re}call(i)}{Anz.Kategorien}$$

Precision gibt die Wahrscheinlichkeit an mit der ein gefundenes Objekt relevant ist. Ist das Verhältnis der richtig Positiven zu allen gefundenen Objekten. (positiver Vorhersagewert, Relevanz).

Recall gibt die Wahrscheinlichkeit an mit der ein relevantes Objekt gefunden wird. Ist das Verhältnis der richtig Positiven zu allen eigentlich relevanten Objekten. (Sensitivität, Empfindlichkeit)

Es ist kaum möglich beide Maße gleichzeitig zu optimieren, denn eine hohe „Precision" erhöht gleichzeitig die Wahrscheinlichkeit seltene Fälle nicht abzudecken. Und bei der Optimierung von „Recall" wird es dazu führen auch nicht relevante Objekte einzubeziehen.

Unterschied, Wann Welche ??

4. Information-Extraction: Was macht das Lexicon & Morphology Modul? Wann ist eher die Verwendung eines Lexicons, wann eines Morphology Moduls angebracht?

Im „Lexicon & Morphology Modul" werden die Tokens (vom „Tokenization Modul") auf ihre Wortart (Verb, Nomen, Adjektiv, …) überprüft. Dies kann durch einfaches Nachsehen in einem Lexikon oder durch ein morphologisches Modul („Part-of-speech Tagger") durchgeführt werden. Das morphologische Modul versucht die Bedeutung und Funktion eines Wortes durch seine Morpheme (kleinster Wortteil) zu ergründen.Die Wichtigste Aufgabe des L&M Moduls ist die richtige Zuordnung von Eigennamen.

Ein Lexikon kann eingesetzt werden wenn alle Varianten eines Wortes leicht aufgelistet werden können (z.B. im Englischen). Wenn hingegen die Komplexität der Sprache diese Auflistung zu einer schwierigen oder gar unlösbaren (z.B. im Deutschen, wo durch Komposition mehrerer Wörter neue Wörter entstehen) Problematik werden lässt muss auf die Morphologie zurückgegriffen werden.

5. Text summarization: Was versteht man unter abstraktiver bzw. extraktiver Summarization?

Extraktive Extraktion: Relevante Sätze oder Paragraphen werden gesucht. Die gefundenen Stellen werden gereiht und in eine Zusammenfassung kopiert. Morphologische Analyse ist nicht notwendig.

Abstraktive Extraktion: Verwendet entweder Vorwissen über die Struktur der verlangten Information („Domain knowlege") und füllt die Information in „Scripts" oder „Templates", welche dann durch NLG-Systeme in zusammenfassende Sätze transformiert wird.
Oder es wird eine Semantische Repräsentation des Dokuments „gebaut" welche dann wieder mit NLG-Systemen in eine leserliche Form gebracht werden kann.

6. QA: Nennen sie die gebräuchlichsten Komponenten eines Question Answering Systems

Analyze Question
Gather information
Distill answers
Sanity Check
Present answers