

Universität Salzburg
Institut für Computerwissenschaften



Seminar aus Informatik SS 2009

Leiter: Univ.-Prof. Dr. Wolfgang Pree

Meaning-Based Computing

Simon Kranzer, Sabine Klausner

{simon.kranzer sabine.klausner} at cosy dot sbg dot ac dot at

08 July 2009

Inhaltsverzeichnis

Inhaltsverzeichnis	i
Abbildungsverzeichnis	ii
Abstract	iii
1 Einleitung	1-1
1.1 Motivation	1-1
1.2 Unstrukturierte Daten	1-1
1.3 Meaning-Based Computing.....	1-2
2 Knowledge Management.....	2-3
2.1 Fragestellungen	2-3
2.2 Konzepte und Techniken.....	2-4
3 Information Retrieval	3-5
3.1 Boolean Retrieval	3-5
3.2 Tolerant Retrieval.....	3-7
4 Text-Mining	4-8
4.1 Was ist Text Mining und deren Einsatzgebiete?	4-8
5 Video-Mining	5-10
6 Resümee	6-13

Abbildungsverzeichnis

Abbildung 3.1: Incidents Matrix	3-6
Abbildung 3.2: Inverted Index	3-7
Abbildung 5.1: Konzept der Gesetzesfindung von Videos	5-10
Abbildung 5.2: Aufbau eines Filmes.....	5-11
Abbildung 5.3: Explosion eines Autos und die sich daraus ergebende Frequenz.....	5-12

Abstract

This seminar-paper deals with Meaning-Based Computing (MBC). It defines MBC as a combination of Knowledge Management (KM) and Information Retrieval (IR) in structured and unstructured data. As a set of these techniques MBC can provide the methods to process the large amount of digital data produced by organisations in our days. KM thereby helps to provide the data to all involved parties. IR-techniques like Text Mining and Video Mining open an unstructured world to KM and its tools to transform this information into searchable content.

1 Einleitung

Im ersten Abschnitt dieser Arbeit werden, für die weiteren Kapitel grundlegende, Begriffe eingeführt und das Thema motiviert. Im zweiten Kapitel werden verschiedene Prozesse und Werkzeuge vorgestellt, die es ermöglichen (Unternehmens)Wissen zu verwalten und bereitzustellen (*Knowledge Management*). Kapitel 3 beschäftigt sich mit der Gewinnung von neuem Wissen aus bestehenden Daten (*Information Retrieval*). Anhand von Beispielen werden zwei Techniken zum Informationsgewinn in den Kapiteln 4 und 5 vorgestellt (*Text- und Video Mining*). Abschließend zieht das letzte Kapitel ein kurzes Resümee zur Arbeit.

1.1 Motivation

Bis zu 80 Prozent¹ der digitalen Daten liegen in nicht strukturierter Form vor. Diese Informationen werden als unstrukturierte Daten bezeichnet (Abschnitt 1.2) und sind für Computersysteme ungleich schwerer zu interpretieren als Daten in strukturierter Form.

Aus dieser Problematik heraus wurden in den letzten Jahren große Anstrengungen aufgebracht um unstrukturierte Informationsquellen nutzen zu können. Die vorliegende Seminararbeit greift einige der dazu geeigneten Verfahren auf und gibt Beispiele zu deren Anwendung.

1.2 Unstrukturierte Daten

Im Gegensatz zu Informationen die in strukturierter Form, etwa in einer relationalen Datenbank, erfasst sind, existiert eine große Anzahl an unstrukturierten Informationsquellen. Beispiele für derartige Datenquellen sind unter Anderen:

- Dokumente
- Emails
- Aufgezeichnete Telefongespräche
- Online Konversationen
- Multimedia Daten

All diese Daten liegen zwar üblicherweise in digitaler Form vor können jedoch nicht mit klassischen Suchalgorithmen und Data Mining-Techniken durchsucht oder analysiert werden. Meaning-Based Computing soll diese Herausforderung bewältigen.

¹ Autonomy At A Glance, http://www.etalp.com/cms/groups/public/documents/collateral/etalk_media_kit.pdf

1.3 *Meaning-Based Computing*

Es ist nicht ganz klar wann und wo der Begriff Meaning-Based Computing (MBC) erstmals verwendet wurde. Forschungen in den Bereichen *Enterprise Search* und Knowledge Management unter der Verwendung von adaptiven Techniken zur Mustererkennung an der Universität von Cambridge², England, führten 1995 zur Gründung der Firma Autonomy Corporation PLC³, die sich selbst als Anbieter von MBC-Produkten bezeichnet. Autonomy definiert MBC wie folgt:

„Meaning-Based Computing is the technology to capture tacit information from unstructured data!”

MBC soll es Computersystemen erlauben Beziehungen zwischen verschiedenartigen Daten zu erfassen und danach komplexe Analysen durchzuführen um in Echtzeit und ohne manuelle Eingriffe relevante Zusammenhänge aufzuzeigen. Dabei nutzt MBC Techniken die weit über traditionelle Schlüsselwortsuchen hinaus die Bedeutung in und von Daten bzw. Informationen zu Tage fördern sollen.

MBC bezeichnet also ein Set an Technologien versteckte Informationen aus unstrukturierten Daten zu extrahieren. Im weiteren Verlauf dieser Arbeit fasst MBC somit alle Werkzeuge zur Suche, Erfassung und Verwaltung von Informationen in sowohl strukturierten als auch unstrukturierten Daten zusammen. Methoden des MBC sind dabei:

- Spracherkennung, Video-/Bilderkennung
- Extrahierung der Bedeutung von Texten
- Implizierte Suche (Eine Suche impliziert auch ein Umfeld, eine Thematik. Relevante Zusatzinformationen werden automatisiert mit einbezogen.)

Diese Methoden bedienen sich dabei Techniken die in die beiden Bereiche Knowledge Management und Information Retrieval geteilt werden können. Im Folgenden werden diese beiden Bereiche nacheinander diskutiert.

² <http://www.cam.ac.uk/>

³ <http://www.autonomy.com/>

2 Knowledge Management

Knowledge Management (KM), die Verwaltung von Wissen, ist definiert als die Summe der Prozesse und Werkzeuge die eine Organisation nutzt um ihr intellektuelles Kapital zu sammeln, analysieren, speichern und zu verteilen⁴. KM bezieht sich dabei auch auf den optimalen Gebrauch von Sachkenntnis und Bedeutung von Wissen in einer Organisation. In diesem Zusammenhang wird jede Art von Wissen, werden alle mögliche Informationsquellen, strukturierte, unstrukturierte, gesprochene, geschriebene und alle anderen verfügbaren, mit einbezogen⁵.

2.1 Fragestellungen

Wie schon zuvor besprochen besteht die Hauptanforderung an KM im Sammeln und Aufbereiten von für eine Organisation relevanten Daten. Das Problem ist dabei die wesentlichen Informationen aus den gesammelten Daten herauszufiltern und in brauchbares Wissen aufzubereiten und von nicht relevantem Informationen zu trennen. Oft kann durch eine Liste von Fragen dargestellt werden wo diese Trennung liegt. Die Antworten auf diese Fragen liefern dabei die gewünschten Informationen.

Anhand eines Beispiels für einen Softwarehersteller soll diese Vorgehensweise im Folgenden dargestellt werden⁶:

- Wissen Sie, warum Ihre Kunden bei Ihnen kaufen?
- Kennt die Mehrzahl der Mitarbeiter die 10 häufigsten Verbesserungswünsche der Kunden?
- Wissen die Mitarbeiter, wo im Unternehmen die Experten für verschiedenste Bereiche sitzen?
- Kommt der Mitarbeiter sich manchmal vor wie ein Goldwäscher, der aus der riesigen Datenflut mühsam die wenigen brauchbaren Körnchen filtern muss?
- Sind die Erfolgsfaktoren der Unternehmung messbar?
- Werden Arbeiten im Unternehmen in mehreren Abteilungen mitunter redundant ausgeführt, da sie voneinander zu wenig wissen?

⁴ Davenport, Tom (2008), "Enterprise 2.0: The New, New Knowledge Management?", Harvard Business Online;

⁵ Thompson, Mark P.A. & Geoff Walsham (2004), "Placing Knowledge Management in Context", Journal of Management Studies 41 (5);

⁶ MICROSOFT TechNet (2001), „Knowledge-Management: Schlagwort oder erfolgskritische IT-Anwendung?“, Mobile Lifestreams: Data on 3G

- Gibt es den kompletten Überblick über Kunden, Projekte, Lieferanten, Wettbewerber?

Anhand dieser Fragen müssen nun Konzepte, Techniken und Methoden definiert werden die zur Beantwortung der Fragestellungen beitragen. Das Ergebnis dieser Ermittlungen liefert dann die Funktionalitäten des KM, das in der jeweiligen Organisation notwendig ist um relevantes Wissen zu erfassen.

2.2 Konzepte und Techniken⁷

Zur Informationsgewinnung stehen dem KM unter Anderen folgende Konzepte und Techniken zur Verfügung⁸:

- Knowledge Maps
- Groupware
- Data- and Text-Mining (Kapitel 4, Kapitel 5)
- Structuring of data
- Workflow
- E-Learning
- Search & Retrieval
- Portals
- Intranet

All diese Methoden können unter dem Begriff Information Retrieval Techniken zusammengefasst werden. Der nächste Abschnitt diskutiert die den gängigen Techniken zu Grunde liegenden Herangehensweisen bei der Gewinnung von Informationen aus digitalen Daten.

⁷ “Wissensakquisition”, http://cartoon.iguw.tuwien.ac.at/fit/fit07/konzept_knowman.html

⁸ Noll, Margit, Fröhlich, Doris (2002), „Knowledge Maps of Knowledge Management Tools”, PAKM 2002

3 Information Retrieval⁹

Wie schon in den vorhergegangenen Kapiteln verdeutlicht ist es nicht ganz einfach die einzelnen Begriffe im Zusammenhang mit MBC konkret einzugrenzen. Auch das Information Retrieval (IR) beinhaltet jede Art von Auffindung von Informationen. Als Forschungsbereich ist es, weniger breit, definiert als:

“Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).”¹⁰

Anhand eines Beispiels soll nun im Folgenden das Prinzip des Findens von Information in umstrukturierten Daten gezeigt werden. Dabei werden die Techniken „*Boolean Retrieval*“ und „*Tolerant Retrieval*“ vorgestellt.

3.1 Boolean Retrieval

Selbst bei relativ einfachen Fragestellungen ist die Suche in unstrukturierten Daten mit klassischen Suchverfahren sehr aufwendig. Wenn man beispielsweise in den digital vorhandenen Werken von Willhelm Shakespeare alle Kapitel, die die Namen „Caesar“ und „Brutus“ enthalten nicht aber den Namen „Calpurnia“ finden wollte, könnte man mit klassischen Methoden wie folgt vorgehen:

1. Durchsuche alle Kapitel nach den Namen „Caesar“ und „Brutus“ und notiere die, in denen beide Namen vorkommen.
2. Durchsuche die notierten Kapitel nach dem Namen „Calpurnia“ und eliminiere die aus der Liste, in denen der Name vorkommt.

Der beschriebene Vorgang könnte etwa mit „grep“ oder ähnlichen Werkzeugen durchgeführt werden. Da der gesamte Text einmal und das Ergebnis noch einmal durchsucht werden müsste, ist diese Herangehensweise insbesondere bei sehr großen Datensätzen nicht optimal. So umfassen die Werke von Shakespeare zwar nur knapp eine Million Wörter und ein modernes Computersystem könnte die Fragestellung aus dem Beispiel auch auf die beschriebene Weise relativ schnell beantworten, betrachtet man aber allein die Webseiten die unter dem Schlagwort „Shakespeare“ in Google gefunden werden können (> 42 Millionen) bedarf es besser geeigneter Techniken um selbst diese simple Aufgabenstellung in vernünftiger Zeit zu lösen.

⁹ Manning, Raghavan, Schütze (2009), „An Introduction to Information Retrieval“, Cambridge University Press

¹⁰ Manning, Raghavan, Schütze (2009), „An Introduction to Information Retrieval“, Cambridge University Press

Eine sehr grundlegende Technik dazu ist das Boolean Retrieval (BR). Dabei wird folgendermaßen Vorgegangen:

1. Bei Aufnahme eines Textes in das System wird einmalig ein lineares Durchsuchen des Textes vorgenommen.
2. Wörter die in dem Text vorkommen werden in einem Index (Incidents Matrix) mit deren Stelle im Dokument erfasst.
3. Dieser Vorgang wird für jedes neue Dokument durchgeführt, es entsteht ein Index über alle Texte im System.

Im Beispiel, der Einfachheit halber nun mit Büchern und nicht mit einzelnen Kapiteln, entsteht dadurch dann eine Matrix wie in Abbildung 3.1. Die Spalten identifizieren dabei die Stücke und die Zeilen die Wörter und deren Vorkommen. Für das Wort „Brutus“ kann aus dieser der Vektor 1 1 0 1 0 0 herausgelesen werden. Analog dazu für „Caesar“ (1 1 0 1 1 1) und 0 1 0 0 0 0 für „Calpurnia“.

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth	...
Antony	1	1	0	0	0	1	
Brutus	1	1	0	1	0	0	
Caesar	1	1	0	1	1	1	
Calpurnia	0	1	0	0	0	0	
Cleopatra	1	0	0	0	0	0	
mercy	1	0	1	1	1	1	
worser	1	0	1	1	1	0	
...							

Abbildung 3.1: Incidents Matrix – Shakespeare aus ¹¹

Die Aufgabenstellung kann nun wie folgt formuliert werden:

$$(110100 \wedge 110111) \wedge \neg(010000) \Leftrightarrow \text{Brutus AND Caesar AND NOT Calpurnia}$$

Dies führt zum Ergebnis:

100100

Die Antwort auf die Frage sind also die Bücher „Antonius und Kleopatra“ und „Hamlet“.

Obwohl auch bei dieser Vorgehensweise jeder Text einmal komplett durchsucht werden muss und zusätzlich Aufbau und Verwaltung des Index realisiert werden müssen wird sofort deutlich, dass dies für beliebig viele und beliebig komplexe Fragestellungen nur einmal pro Dokument notwendig ist.

Es existieren noch einige Verbesserungen des BR, dazu sei jedoch auf die diesem Kapitel zugrunde liegende Literatur verwiesen.

¹¹ Manning, Raghavan, Schütze (2009), „An Introduction to Information Retrieval“, Cambridge University Press

3.2 Tolerant Retrieval

Die im vorhergehenden Abschnitt besprochene Technik (BR) ist sehr effizient für eindeutige Fragestellungen, die als logische Ausdrücke formuliert werden können. Sehr oft ist dies nicht ohne weiteres möglich. Rechtschreibfehler, alternative Schreibweisen und Synonyme sind nur einige Beispiele für den Bedarf nach unscharfen Suchverfahren. Tolerant Retrieval (TR) erlaubt diese Unschärfen, diese Technik wird im vorliegenden Abschnitt genauer vorgestellt. Ähnlich der Matrix aus Abschnitt 3.1 kann ein Dokument auch auf folgende Weise indiziert werden (Abbildung 3.2):

Brutus	→	1	2	4	11	31	45	173	174
Caesar	→	1	2	4	5	6	16	57	132 ...
Calpurnia	→	2	31	54	101				

Abbildung 3.2: Inverted Index – Shakespeare aus ¹²

Abbildung 3.2 zeigt einen Index (Inverted Index) in dem jedes Wort und dessen Vorkommen im Text gespeichert wird. Im Gegensatz zum früheren Ansatz muss hier jedoch nicht für jedes Vorkommen eine neue Spalte hinzugefügt werden. Die Liste der Wörter wird dabei als Dictionary und die Vektoren werden als Postings bezeichnet. Mit Hilfe von binären Bäumen können die Terme im Dictionary in kleinere Einheiten unterteilt werden (etwa Buchstaben) und somit in Klassen unterteilt werden. Damit und mittels Wildcards können auch Suchanfragen wie etwa „Calpurnia“ oder „Klapurnia“ als „*alpurnia“ formuliert werden. Ebenso wie beim BR existieren auch beim TR erweiterte Methoden, die weitere unscharfe Suchen ermöglichen.

In den nachfolgenden Kapiteln werden Text- und Video Mining als konkrete Repräsentanten von IR vorgestellt.

¹² Manning, Raghavan, Schütze (2009), „An Introduction to Information Retrieval“, Cambridge University Press

4 Text-Mining

4.1 Was ist Text Mining und deren Einsatzgebiete?

Text-Mining ist ein relativ neues Verfahren, dass sich mit der inhaltlichen Auffassungsgabe und der Relevanz von Texten beschäftigt. Gerade zur Zeit der elektronischen Datensammlung sei es in einfachen Textdokumenten oder in Datenbanken und weiters der Nutzung von online erreichbaren Informationsplattformen gewinnt das Thema einer effektiven und effizienten Datenerfassung vermehrt an Bedeutung.

Bei Text-Mining handelt es sich um die Aufgabe Informationen nach Richtigkeit und Wichtigkeit einzuordnen, je nachdem wozu diese verwendet werden sollen.

Sowohl im privaten Haushalt als auch in Unternehmen und Regierungen werden über zwei Drittel bereits in digitaler Form erfasst. Dieser sogenannte Informations-Overload stellt für den Leser eine Herausforderung dar, da bei der Suche nach Daten diese dementsprechend zu Filtern sind.

Um dieses Problem nun unter Kontrolle zu bekommen wurden diverse Technologien und Prozessschritte definiert, die in Abbildung 4.1 zu sehen sind.

Bei den Prozessschritten wird zwischen der Suche, Vorverarbeitung, Bewertung und Selektion und der Extrahierung und Mustererkennung unterschieden. Jeder Prozessschritt wird durch mindestens eine Technologie wie Information Retrieval, Computer Linguistik, Klassifizierung, Clustering und/oder Feature Extraktion unterstützt.

Jeder Prozessschritt setzt auf den vorhergehenden auf. Bei der Suche wird durch beispielsweise Suchmaschinen zwischen relevanten und irrelevanten Texten unterschieden. Information Retrieval ist eine Methode die sich mit der Organisation und dem Auffinden von Informationen beschäftigt. Dabei wird auf die Suche nach speziellen Themenbereichen acht genommen. Hilfestellung dafür bieten das Vektorraummodell und das Latence Semantic Indexing. Kurz beschrieben handelt es sich beim ersten Ansatz um die Darstellung von Dokumenten, die anhand von einem Vektorraum darstellbar wird und beim zweiten um das Finden von Termen durch die Wahrscheinlichkeit des Auftretens. Dabei wird ein sogenannter Cluster gebildet der die Suche auf diese Terme einschränkt. Mehr Details dazu in Witte¹³.

¹³ Witte, Mülle (2006), „Wissensgewinnung aus natürlichsprachigen Dokumenten“, Universität Karlsruhe, Bericht 2006-5

Nachdem die Suche alleine noch zu viele Datenmengen aus Dokumenten enthält wird in Folge der Ansatz der Vorverarbeitung eingesetzt. Es werden die Computerlinguistik, die statistische Sprachverarbeitung und die Verwendung von Makrostrukturen in Texten als unterstützende Technologie eingesetzt. Bei der Methode der Computerlinguistik wird auf Wort, Sätze und Phrasen acht genommen und nicht nur auf diverse Satzstrukturen. Die ist möglich, da gewisse Konstruktmuster aus dem Zusammenhang heraus erkennbar sind. Bei der statistischen Sprachverarbeitung wird mit Schlüsselbegriffen in Texten (Wichtigkeit eines Wortes in Texten) oder Phrasenmustern gearbeitet.

Bei den Makrostrukturen wird der Text nach Interessensgebieten des Lesers strukturiert. Die Suchinteressen werden gefiltert und bei der Selektion von Texten berücksichtigt. Beim Prozessschritt Bewertung und Selektion werden die zuvor gesammelten Texte je nach Themengebiet in Gruppen eingeteilt. Hilfestellung bieten die Klassifizierung für die Gruppierung nach Themengebieten und das Clustering für das rasche Auffinden von gleichartigen Dokumenten und Texten ohne vordefinierte Terme. Bei beiden Arten werden wieder spezielle Methoden angewendet die aus [] entnommen werden können.

Beim letzten Prozessschritt, nämlich der Extrahierung und Mustererkennung handelt es sich um die Suche nach Sequenzen im Text. Bei der Technologie des Data Minings handelt es sich um das Auffinden von Informationen die aus anderen resultieren.

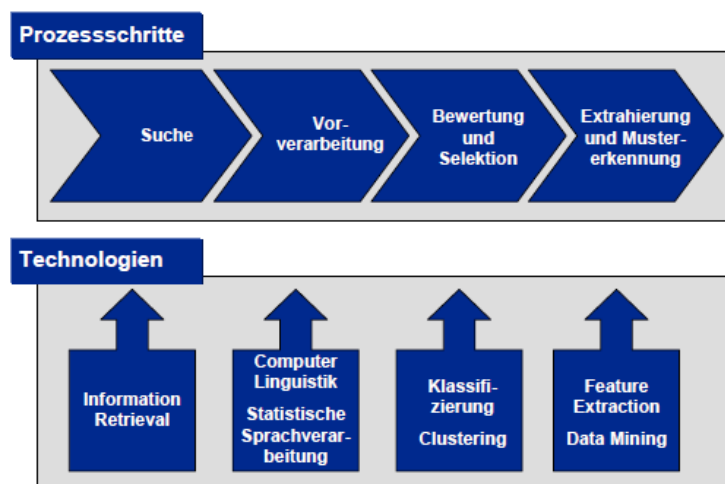


Abbildung 4.1: Text Mining dargestellt in deren Technologien und deren Prozessschritte

Der Einsatz von Text Mining findet bei der Analyse von Texten - Kommunikationsüberwachung bei der Aufklärung eines Verbrechens durch beispielsweise Suche nach Hinweise in E-Mails oder SMS und so weiter - an Bedeutung.

5 Video-Mining

Wie auch beim Text Mining werden Videos durch die Digitalisierung dieser immer häufiger in Datenbanken erfasst. Die Analyse der Daten erfolgt durch Indizierung (Suche auf Basis von Schlüsselbegriffen). Beim Video Mining handelt es sich um ein sogenanntes "Rich Medium", da nicht nur Text sondern Sprache, Bilder, Musik und Eindrücke des Zuschauers ein Video ausmachen und beschreiben.

Die Umsetzung von Video Mining erfolgt durch Erforschung von Konzepten und das Auffinden der Bedeutung eines Videos. Weiters wird die sogenannte Filmgrammatik in Betracht gezogen. Dabei handelt es sich um Gesetze die in Videos beachtet werden, damit der Charakter dessen für den Zuschauer erkennbar wird.

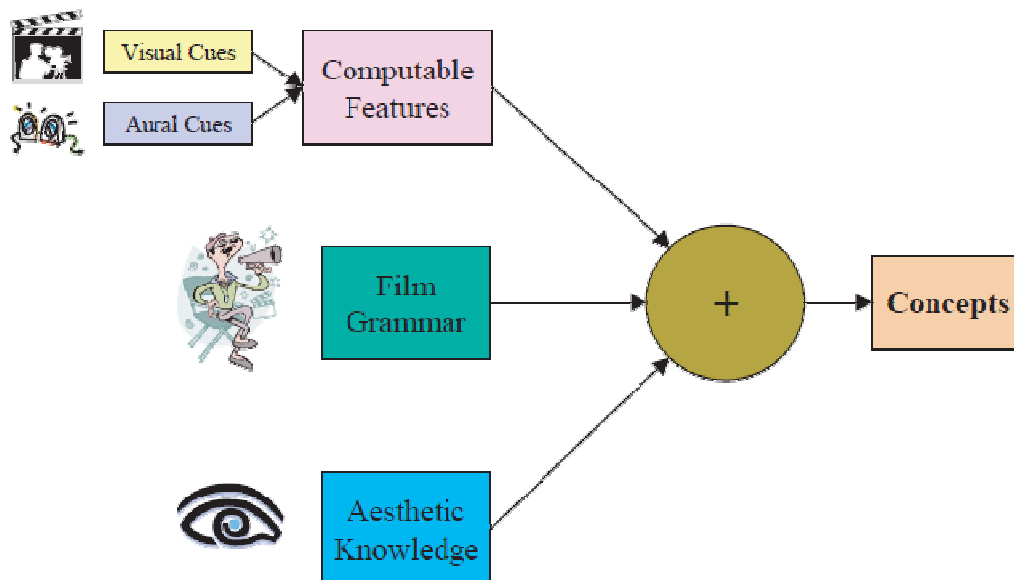


Abbildung 5.1: Konzept der Gesetzesfindung von Videos

In Abbildung 5.1 wird der Ansatz für die Umsetzung des Konzeptes eines Videos gemacht. Dabei werden sowohl visuelle als auch hörende Merkmale die eine Zuseher wahrnimmt gesammelt. Die beiden Informationen bilden die sogenannten "berechenbaren Eigenschaften". Weiters werden die Filmgrammatik, also welche Sprache, Sprach-, Farb- und Musikeffekte zu welcher Szene folgen erfasst. Diese Grammatik leitet sich aus der Handlung ab, beispielsweise weisen Talkshows andere Merkmale als ein Kriminalfilm auf. Weiters werden die ästhetischen Eindrücke des Zusehers für die Erstellung des Konzeptes zusammengefasst. Alle diese Eigenschaften zusammen bilden das Konzept eines Filmes/Videos.

Weiters wird in Abbildung 5.2 beschrieben in welche Schichten ein Video unterteilt ist. Die kleinste Einheit eines Videos bildet das Frame, dass einem Buchstaben im Alphabet gleicht. Weiters werden mehrere Frames in Shots (visuelle Eindrücke) zusammengefasst, die ein Wort bestehend aus mehreren Buchstaben bilden. Mehrere Shots (aneinandergereihte Bilder) erzeugen einen sogenannten Beatstream. Eine bestimmte Konstellation an Shots bilden einzelne Scenes (Textfragmente) die in Folge dessen das Video, die Geschichte bestehend aus Sätzen, ausmachen.

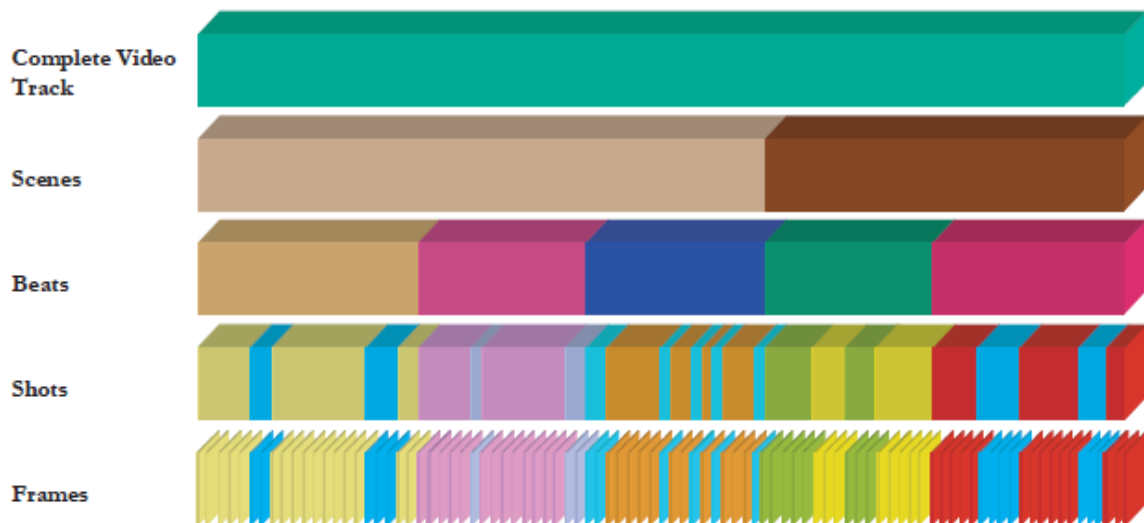


Abbildung 5.2: Aufbau eines Filmes

Die folgende Abbildung 5.3 zeigt eine Möglichkeit der Selektion von Videos durch Aufzeichnung der Geräusche. Dabei wird die Amplitude zum jeweiligen Frame erfasst. Befindet sich die Amplitude bei einer vordefinierten Frequenz kann daraus geschlossen werden, dass es sich um eine bestimmte Szene handelt. Im konkreten Beispiel wird der Brand eines Fahrzeuges und infolgedessen deren Explosion erfasst. Bevor die Explosion erfolgt befindet sich der Geräuschpegel im unteren Frequenzbereich. Die Explosion selber, also der Knall befördert die Amplitude nach oben und hält den Geräuschzustand konstant nach oben.



Abbildung 5.3: Explosion eines Autos und die sich daraus ergebende Frequenz

Daraus kann geschlossen werden, dass es sich um eine Explosion und eventuell um eine Krimi handelt.

6 Resümee

Aufgrund der steigenden Zahl an Daten die in unstrukturierter Form vorliegen ist die Bedeutung von MBC in den letzten Jahren stark gewachsen. Mit Hilfe von Techniken des IR können diese Informationen erfasst und in eine strukturierte Form übergeführt werden. Die relativ komfortable Suche in einer relationalen Datenbank kann dann mehr Wissen aus den gewonnenen Informationen generieren. Unterstützt durch Werkzeuge zum KM können diese wiederum allen Beteiligten Personen einer Organisation in aufbereiteter Form zur Verfügung gestellt werden.