Automatische Klassifikation

Quelle:http://www.isn-oldenburg.de/images/products/medium/classifier.jpg

Themenübersicht

- ✓ Definition
- ✓ Einsatzgebiete
- ✓ Selber ausprobieren
- ✓ Verfahren
  - ✓ Statische Verfahren
  - ✓ Lernende Verfahren

→ Definition (automatische) Klassifikation

= Zuordnung von Dokumenten zu **bereits vorher festgelegten Klassen**

Zielsetzung:

✓Dem Nutzer Informationen über inhaltlich ähnliche Dokumente geben
✓Hierarchisches Browsing in der Ergebnismenge ermöglichen

Einsatzgebiete für automatische Klassifikation

Industrie: Verwaltung von Ersatzteilen

simussystems

Social Network

Monitoring radian6

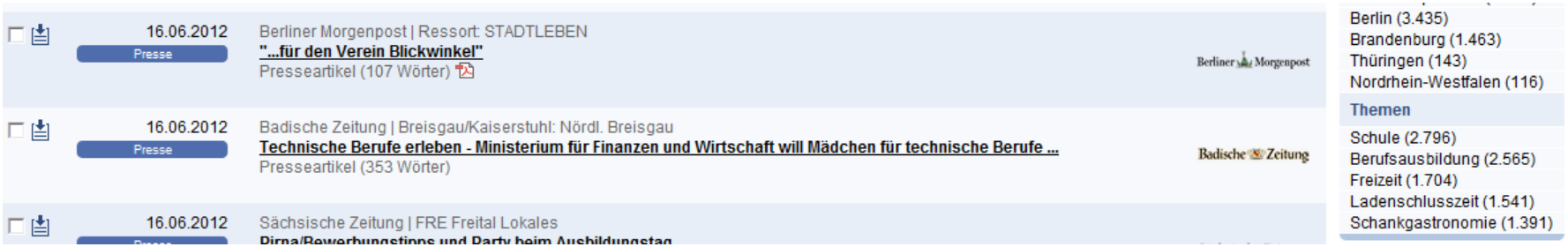
Bibliotheks**kataloge** zur Verbesserung der Navigierbarkeit von Trefferlisten

BASE
Bielefeld Academic Search Engine

Viele Softwareanbieter für:
Enterprise Content Management
Dokumentenmanagement

Lingua Systems
NATURAL LANGUAGE PROCESSING

Pressearchive, Pressedatenbanken – WISO; GENIOS



→ Zwei Verfahren

1. Statische Verfahren: Einfacher **Merkmalsabgleich** zwischen Klassifikationssystem (Terme in der **Klassenbenennung** und in den Benennungen der Unterklasse) und **Dokument.** Diese Verfahren basieren auf dem **Vergleich von Vektoren.**
2. Dynamische Verfahren: ‚**Lernende** Verfahren' der automatischen Klassifikation anhand von **Trainingsdokumenten**

➔üblich ist eine Kombination beider Verfahren

Beispiel: Spamerkennung

Wie kann ein Programm diese Mails automatisch in den Spamordner einsortieren?

Mail a

Sehr geehrte Frau Spree,

Hiermit reiche ich Ihnen meine Hausarbeit über das Referat zu wissenschaftlichen Erkenntnissen zur Wirkungsweise von Viagra nach.

Mail b

Hi,

Today we announce the selling of our new Viagra pills that allow you to enjoy Sex again. Order now

→ Anwendungsbeispiel: Spamerkennung - Merkmalsbestimmung

Klasse/ Attribut	SPAM	Kein Spam
	Viagra	Sehr geehrte
	Sex	Hochachtungsvoll
	selling	wissenschaftlich
	buy	Hausarbeit
	pills	Referat

Mail a

Sehr geehrte Frau Spree,

Hiermit reiche ich Ihnen meine **Hausarbeit** über das **Referat** zu **wissenschaftlichen** Erkenntnissen zur Wirkungsweise von **Viagra** nach.

Mail b

Hi,

Today we announce the **selling** of our new **Viagra pills** that allow you to enjoy **Sex** again.

→ Ähnlichkeitsabgleich der Hinweisterme für Spam mit den Mails

SPAM		a	b	Kein SPAM		a	b
Viagra	1	**1**	1	Sehr geehrte	1	1	0
sex	1	0	1	Hochachtungsvoll	1	0	0
selling	1	0	1	wissenschaftliche	1	1	0
buy	1	0	0	Hausarbeit	1	1	0
pills	1	0	1	Referat	1	1	0
Skalarprodukt: Klasse/Mail		**1**	**4**			4	0

Mail a ist :

= 20% (1/5) Spam

= 80% (4/5) kein Spam

→ **Einordnung kein Spam**

Mail b:

= 0% (0) kein Spam

= 80% (4/5) Spam

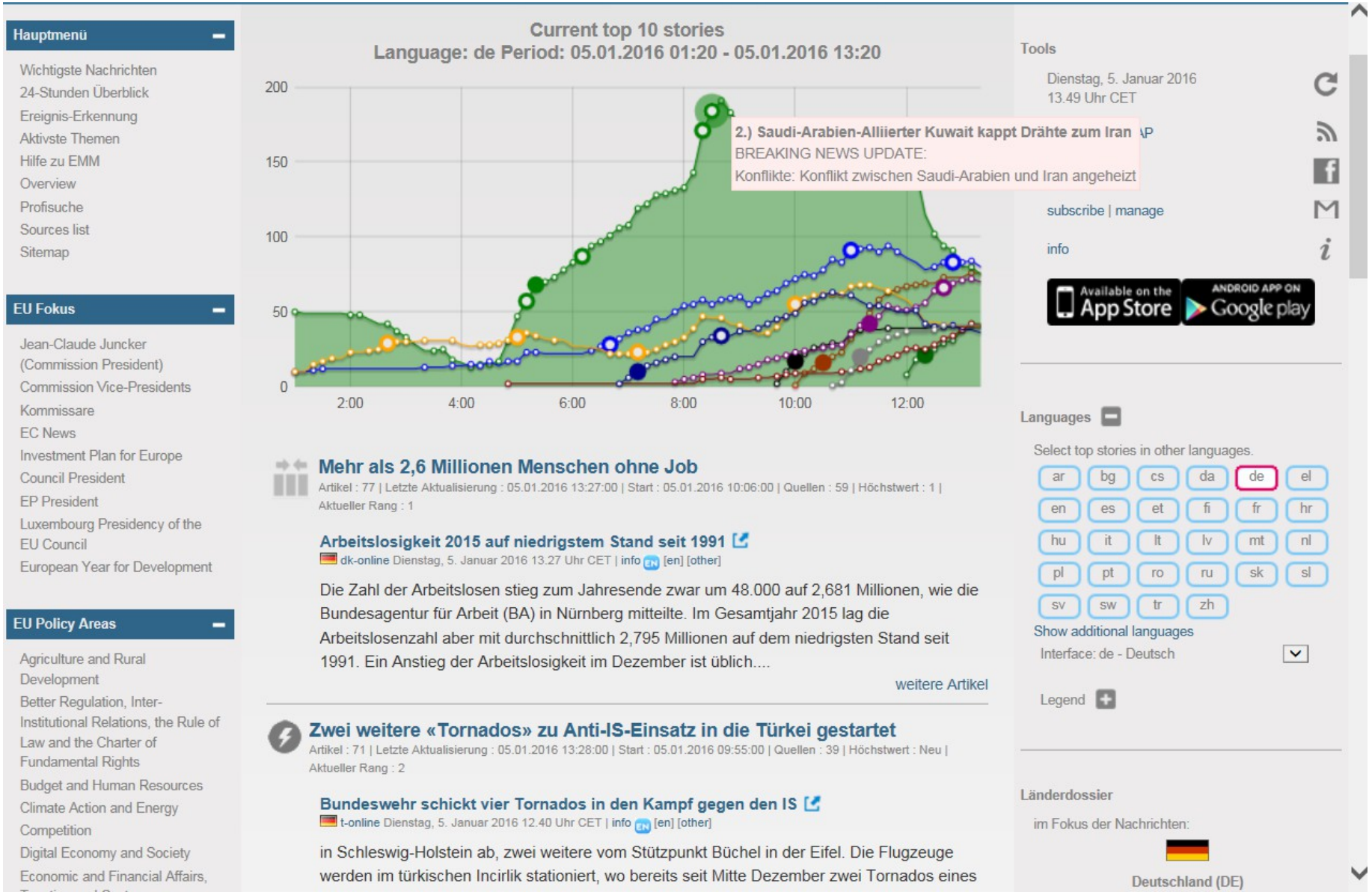
→**Einordnung Spam**

→ 1) Statistische Verfahren : Praktische Nutzung des Verfahrens?

Einfache **Spamfilder** funktionieren nach diesem Prinzip, indem sie einfache Regeln festlegen nach dem Muster:

Wenn die Wörter Viagra, Sex, Werbung, Chance, ….vorkommen, **dann** flagge Dokument als Spam

Beispiel Mediamonintoring NewsBrief



Screenshot: http://emm.newsbrief.eu/NewsBrief/countryedition/en/DE.html

Und jetzt Sie …. (Le 10, Step 2)

Probieren Sie die Mediamonitoring-Anwendungen der Europäischen Kommission News Brief aus:

http://emm.newsbrief.eu/NewsBrief/countryedition/en/DE.html

1) Was sind die wichtigsten News Stories in Deutschland?
2) Was sind die wichtigsten News Stories in Großbritannien? Sie müssen die URL wechseln (http://emm.newsbrief.eu/NewsBrief/countryedition/en/GB.html)
3) Stellen Sie Vermutungen darüber an, wie NewsBrief die Top Stories ermittelt.
4) In der linken Spalte (facettierte Anzeige) können Sie sich Nachrichten zu bestimmten Kategorien anzeigen lassen. Schauen Sie sich die Ergebnisse in der Kategorie EU Policy Areas → Digital Economy and Society und Kriminalität an.
5) Welche Arbeitsschritte sind für eine automatische Zuordnung von Dokumenten zu einer Klassifikation wohl nötig?
    a. Wie müsste die eine vorhandene Klassifikation wie z. B. IPTC aufbereitet werden, damit sie sich für den Einsatz von automatischen Verfahren eignet?
    b. Wie müssten die Artikel aufbereitet werden?
    c. Welche Schwierigkeiten vermuten Sie bei der automatischen Klassifikation? (Belegen Sie Ihre Vermutung am Beispiel)

Every ten minutes and in each of the languages, both applications cluster the latest news items (four hour window or more, depending on the number of recent articles) and present the largest cluster as the current top-ranking media theme (*Top Stories*). The title of the cluster's medoid (the article closest to the cluster centroid) is selected as the most representative title and thus as the title for the cluster. All current clusters are compared to all the clusters produced in the previous round. If at least 10% of the articles overlap between a new cluster and any of the previous ones, the clusters get linked and those articles that have fallen out of the current 4-hour-window get attached to the current cluster. The public web pages are updated every the minutes so that users always see the latest news of the fastest news providers.

Larger new clusters (without overlap to previous clusters) and clusters of a rapidly rising size get automatically classified as *breaking news* so that subscribed users will be notified by email. The statistical breaking news detection algorithm makes use of information on the number of articles and of the number of different news sources, comparing the news of the last 30 minutes with longer periods of time.

Clustering

Cluster benennen

Steinberger, Ralf; Bruno Pouliquen & Erik van der Goot: An introduction to the Europe Media Monitor family of applications. In: Gey, Frederic; Kando, Noriko; Karlgren, Jussi: Proceedings of the SIGIR 2009 Workshop July 23, 2009 Boston, Massachusetts USA : Information Access in a Multilingual World. Boston 2009, S. 1-9. Online: https://www.researchgate.net/profile/Erik_Van_der_Goot/publications. Zugriff: 2016-01-05

All news items are additionally categorized into hundreds of categories. Categories include geographic regions such as each country of the world, organizations, themes such as *natural disasters* or *security*, and more specific classes such as *earthquake*, *terrorism* or *tuberculosis*. Articles fall into a given category if they satisfy the category definition, which consists of Boolean operators with optional vicinity operators and wild cards. Alternatively, cumulative positive or negative weights and a threshold can be used. Uppercase letters in the category definition only match uppercase words, while lowercase words in the definition match both uppercase and lowercase words. Many categories are defined with input from the institutional users themselves.

Zuordnung zu Kategorien/Klassen

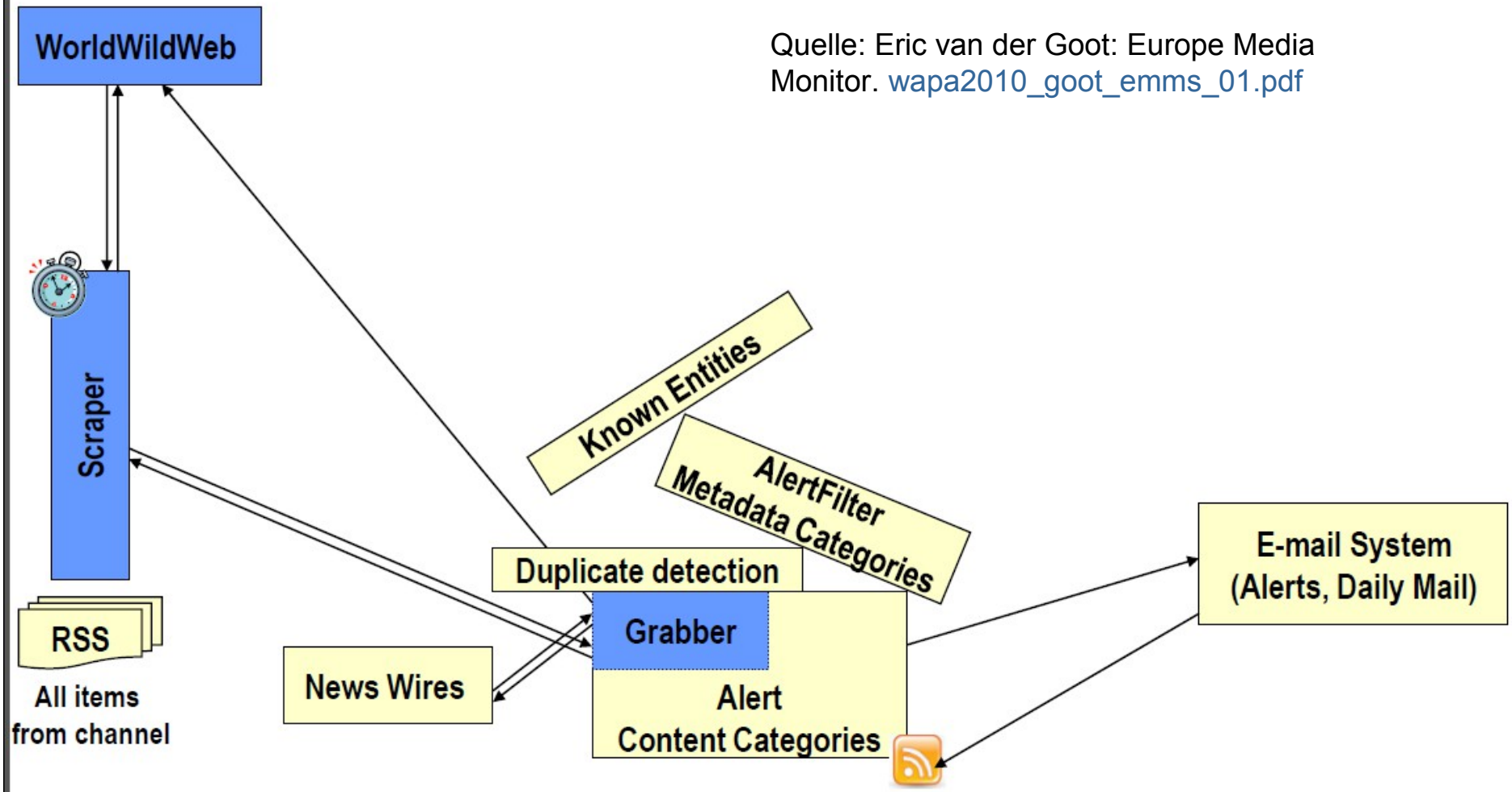
Definition der Klassen:

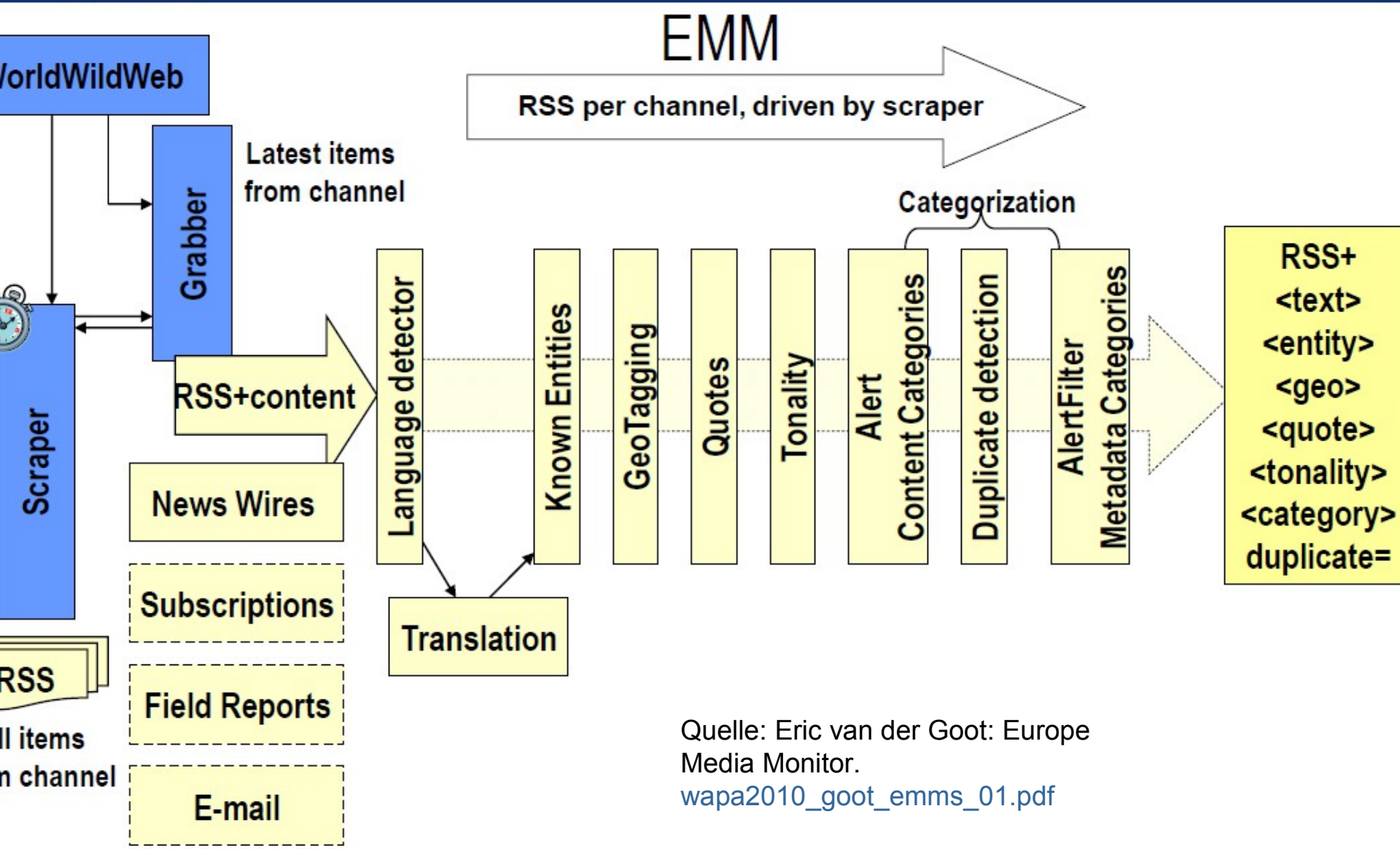
als komplexe Suchstrings (reguläre Ausdrücke)

Steinberger, Ralf; Bruno Pouliquen & Erik van der Goot: An introduction to the Europe Media Monitor family of applications. In: Gey, Frederic; Kando, Noriko; Karlgren, Jussi: Proceedings of the SIGIR 2009 Workshop July 23, 2009 Boston, Massachusetts USA : Information Access in a Multilingual World. Boston 2009, S. 1-9. Online: https://www.researchgate.net/profile/Erik_Van_der_Goot/publications. Zugriff: 2016-01-05

Development history

Quelle: Eric van der Goot: Europe Media Monitor. wapa2010_goot_emms_01.pdf

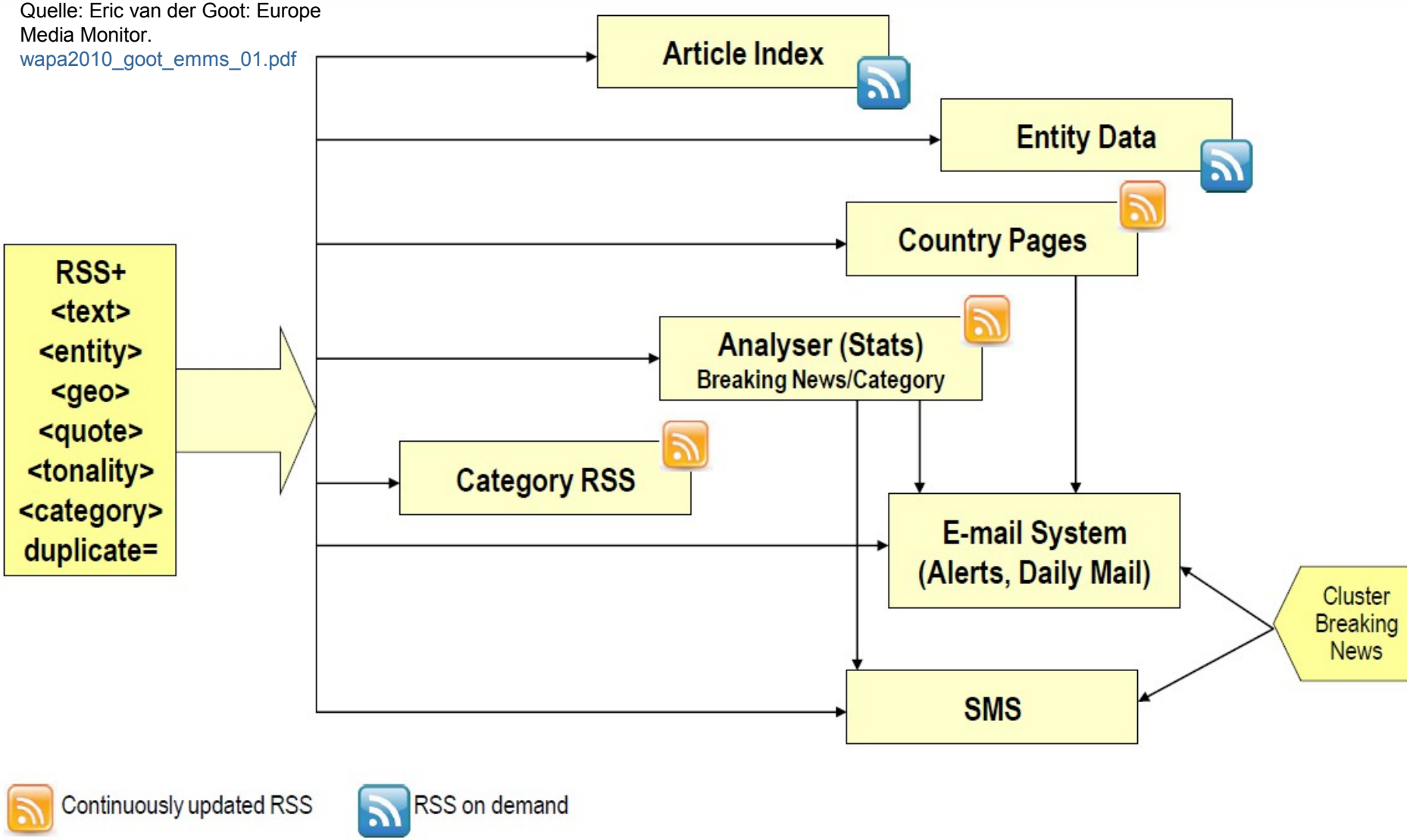




Quelle: Eric van der Goot: Europe Media Monitor. wapa2010_goot_emms_01.pdf

Item based information systems

Quelle: Eric van der Goot: Europe Media Monitor.
wapa2010_goot_emms_01.pdf



→ 1) Statistische Verfahren : Praktische Nutzung des Verfahrens?

Lexis-Nexis wendet ein solches Verfahren zur Einordnung von Pressetexten in 900 Themenklassen an.

SCOPE NOTE: JOINT VENTURES targets joint partnerships between companies where a third, new company is formed and co-owned by the partners. The scope includes announcements of new ventures, failed ventures and joint venture banks.

THRESHOLD = 4 {minimum score needed to tag a document}

HEADLINE COUNT = 2 {score given to a phrase matching in the headline segment}

BODY COUNT = 1 {score given to a phrase matching in the body segment}

VERY STRONG TERMS	STRONG TERMS	STRONG RELATED TERMS	WEAK TERMS	NEGATIVELY WEIGHTED TERMS
"concentrative" j/v	j-venture	form a venture	% of the arrangement	high school jv
"concentrative" j/v	j-ventures	form the venture	% ownership	joint vision 2010
"concentrative" joint venture	j/venture	form ventures	% stake	junior varsity
"concentrative" joint ventures	j/ventures	formation of a venture	business combination	jv 2010
approve the j/v	joint venture	formation of the venture	business combinations	jv boys
approve the j/v	joint venturer		agreement	jv girls
approve the joint venture	joint venturers		agreements	jv squad
to form joint venture	joint ventures		form joint	
to form joint ventures	joint-venture		form jointly	
	joint-venturer			
	joint-venturers			
	joint-ventures			

Quelle: Barbara Quint: Lexis-Nexis Smart Indexing Technology.
http://newsbreaks.infotoday.com/nbreader.asp?ArticleID=17877

→ 1) Statische Verfahren: Beispiel Dokumentklassifikation Bibliothek

1. Merkmale eine Klassifikation werden festgelegt
   - Merkmale können einer Klasse zugeordnete Wörter sein.
   - In der Universalklassifikation Dewey Decimal Classification (DDC) wird die Klasse **Informatik** durch Benennungen der Unterklassen: Wissen, Buch, Systeme, Datenverarbeitung, Computerprogrammierung, Programme, Daten, …definiert
2. Das Vor- oder nicht Vorkommen bestimmter Terme im Dokument wird mit dem Vorkommen dieser Terme in der Merkmals-Beschreibung der Klassen verglichen
3. Dies kann über die Bildung von Skalarprodukten zwischen den Dokumentvektoren (bestimmt über die Deskriptoren) und den Vektoren der Klassen (bestimmt über die Klassenbeschreibungen) geschehen

Beispiel: Zuordnung von Webdokumenten zu einer Universalklassifikation:
http://act-dl.base-search.net/textclassifier
Zu klassifizierender Beispieltext:
http://www.haw-hamburg.de/fakultaeten-und-departments/dmi.html

Throwing out LifeLines

Shaun knows he's dyslexic – at 23 he's young enough to have been "statemented" at school, but like many other children struggling to keep up, the help arrived to little too late. Shaun was already wagging school at 12, and between his own withdrawal and threats of exclusion from the school, it wasn't until he pitched up in a young offender's institution that his learning needs were assessed. Yes, he was diagnosed as dyslexic -again, something he already knew, and although he received support and learning in prison, the sentence was short, and he didn't manage to read his way to the end of it.

Working as the Writer in Residence at HMP Manchester I get to see first hand the daily emergence into a world of literacy of men who thought they would never read for pleasure, never pen a poem, never ask for a dictionary, never hold a library card. This year I've worked alongside tutors, uniformed staff and librarians who work together to make sure someone like Shaun has some choices. Time, attention, care, support, professionalism are the lifelines thrown out to prisoners and when they are caught by men like Shaun the outcomes can be literally life changing.

Poetry or Crime

I'm based in the prison Library, where tides of men arrive, sometimes eager and focused to use their 20 minutes amid the well stocked shelves, sometimes unsure of what a Library for, but keen for a change of scene from the wings and a chance to get out of their cells. The Librarians, Orderlies and Library Officer answer a battery of requests "Love Poetry? That shelf on the right. True Crime ? Over there. Starting up a Business? I can look for you. Archbold's Criminal Pleading – Evidence and Practise? Reference only, but mark the pages and we'll photocopy them for you. Daily Mirror? It's out at the moment…" And then there's less specific questions, by new readers, emerging readers, returning readers, unsure readers:
"I want something to read." "Something to read?" "Yeah. Something to read." But what?

Seven years on from Young Offenders, Shaun's on remand, this time in an adult jail. He's been working as a plasterer, supporting the two children he adores and is missing terribly. He knows he made a stupid choice and a big mistake that night. Now he's with a group of men sitting an assessment test to gauge their literacy and numeracy skills. The room grows quiet as men of all different ages, nationalities, attitudes begin the unfamiliar task of choosing the right verb to fill a gap in a sentence. Shaun begins filling in his name, then he looks around at the others bent over the test. He looks over at the Advice and Guidance worker. She's skilled, warm and funny, and she's already explained that she's here if anyone wants to talk anything through. He comes over with his paper, his first name written in a spidery hand. "Miss, I know I'm dyslexic. They told me at school. I have tried to do something about it, but I've been working and I've got two little kids. My daughter's started school and she loves it, coming back with reading books. I want to be able to help her. I know I've got to do something about it and I might as well have a go while I'm in here"

The AIG worker sets out some choices – there's the Toe by Toe programme, where trained prisoners offer peer support on the wings, there's an intensive Skills for Life class with a great success rate and one to one help from volunteers, or there's structured support in the workshops from Basic Skills tutors. Shaun chooses the Skills for Life class; he wants to learn now and he wants to learn fast. Good choice. He's got to get something good out of this bad situation.

Something to Read

When prisoners like Shaun want "Something to read" what will they choose? Its heartening to see the growing range of books designed specifically with adult beginner and emergent readers on the shelves, but adult readers are a diverse group of people with a diverse set of experiences. What sort of books could a new reader in prison find that would reflect their experiences or even encourage them to write about their own lives? It was with this question in mind that I developed the LifeLines Prison Writing project for the Writers in Prison Network and the Indigo Trust. Evolving from a previous Writers in Prison project called

… welcher Klasse ordnet das System wohl diesen Text über ein Creative-Writing-Projekt in einer Gefängnisbiblioth ek zu?

Quelle: Amanda Wait: Throwing out LiveLines. http://www.newleafbooks.org.u

ACT-DL
The automatic classification toolbox for Digital Libraries

Start | Text Categorizer | Web Categorizer | PDF Categorizer | API | Publica

Categorization Results

Below are the results for your input

DDC Level 1

DDC	Confidence
360: Social problems & services; associations	0.96
410: Linguistics	0.04
027: General libraries	0.00
025: Library operations	0.00
020: Library & information sciences	0.00
780: Music	0.00
370: Education	0.00
306: Culture & institutions	0.00
120: Epistemology, causation & humankind	0.00

… nicht schlecht …

Sreenshot ADC-CL: DDC-Classifier Ergebnis

→ 2) Dynamische Verfahren : Optimierung durch Lernen

- ✓ Voraussetzung:

  Bestand von Trainingsdokumenten, die intellektuell Klassen zugeordnet wurden, ist vorhanden
- ✓ Vorgehen
  - ✓ Analyse der Trainingsdokumente
  - ✓ Ermittlung der Eigenschaften der Dokumente, die bereits einer Klasse zugeteilt wurden

    Eigenschaften sind z. B. das Vorkommen und Gewicht bestimmter Indextermini **in den Dokumenten**
  - ✓ Berechnung der Wahrscheinlichkeit, dass ein bestimmtes Dokument, in dem das Wort x vorkommt, der Klasse y zugeordnet wird

→ 2) Dynamische Verfahren : Erlernen der Klassenzugehörigkeit

Formel **Naiver Bayes Algorithmus**:

Anzahl der Dokumente mit Wort x im Trainingsbestand, **die Klasse y zugeteilt** sind
Anzahl der Dokumente mit Wort x, **die nicht y zugeteilt** sind

Im Trainingsbestand sind 8 Dokumente mit Wort „Viagra “ der **Klasse Spam** zugeteilt.
Viagra kommt in 10 Emails vor, **die kein Spam** sind.

= 8/10 -> 0,8
→Die Wahrscheinlichkeit, dass ein Dokument mit „Viagra “ der **Klasse Spam** zugeordnet wird, liegt bei 0,8

Im Trainingsbestand sind 2 Dokumente mit „Kuss“der **Klasse Spam** zugeteilt.
„Kuss“ kommt in 50 Dokumenten vor, die **kein Spam** sind.

=2/50 ->0,04
→Die Wahrscheinlichkeit, dass ein Dokument mit „Kuss “ der **Klasse Spam** zugeteilt wird, liegt nur bei 0,04

Beispiele für Kategorisierungssoftware

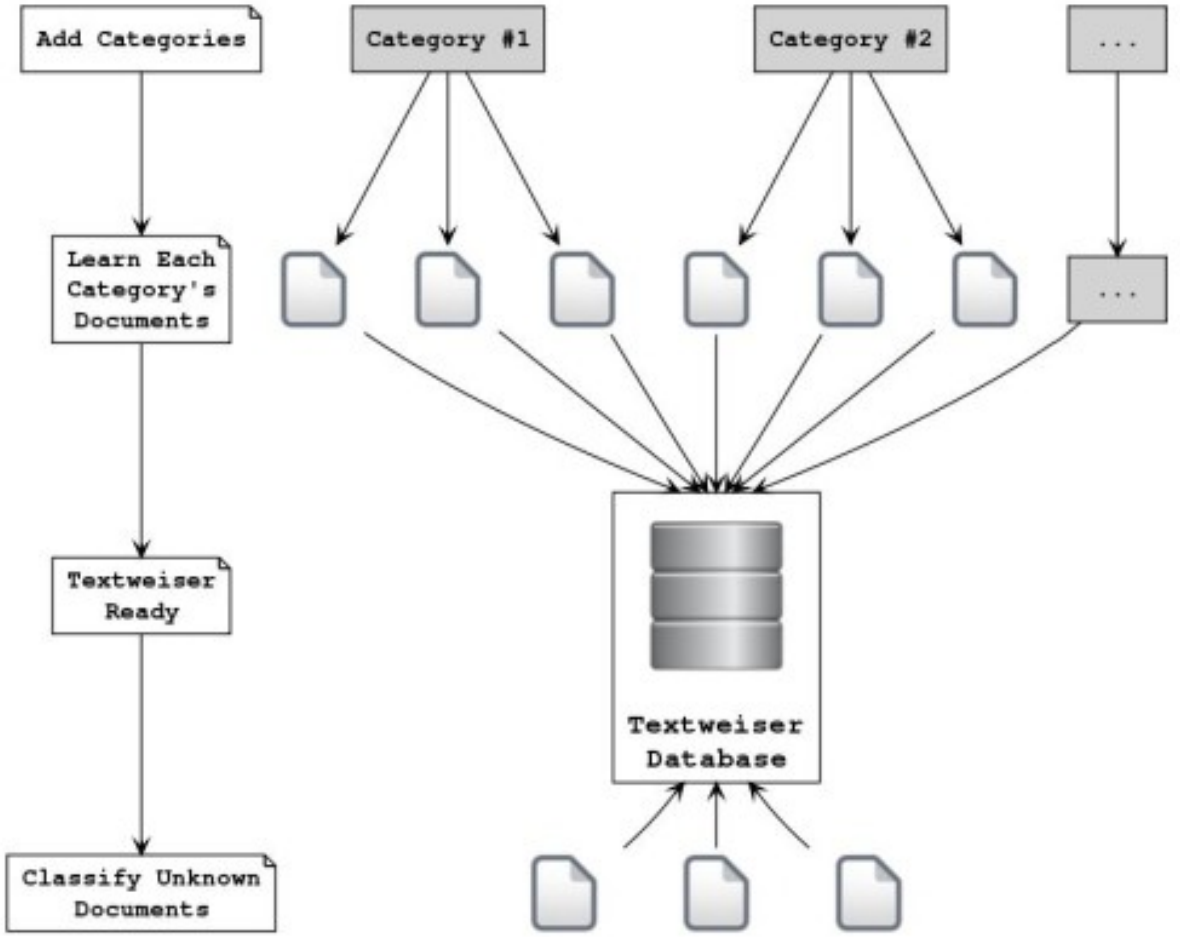
Radian6 – Posts in sozialen Netzwerken kategorisieren

http://www.youtube.com/watch?v=QTqGSLjeqWM. Abruf: 2013-05-12

Textweiser

http://www.lingua-systems.de/text-classifier/textweiser-library/video-demo.html
. Abruf: 2013-05-12

Visualisierung lernende Verfahren automatische Klassifikation



Ablauf: Textweiser
http://www.lingua-systems.de/text-classifier/textweiser-library/workflow.html