Verteidigung der Diplomarbeit

Nachrichtenklassifikation als Komponente in WEBIS

Björn Krellner

<bjk@informatik.tu-chemnitz.de>

TECHNISCHE UNIVERSITÄT CHEMNITZ

Fakultät für Informatik

Professur Informationssysteme und Softwaretechnologie

28.09.2006

Inhalt

Informationsverarbeitung

- Daten $\neq$ Informationen
- Daten:
  - Angaben über verschiedenste Fakten oder Zusammenhänge
  - Objektiv wahrnehmbar und potentiell verwertbar
- Informationen:
  - Erkennbare Semantik
  - Beseitigung einer Ungewissheit
- Finanzwelt:
  - Einzelner Aktienkurswert: *Datum*
  - Kurswert einer Firma zu mehreren Zeitpunkten: *Information* bei entsprechender Fachkenntnis

Informationsverarbeitung

- Daten $\neq$ Informationen
- Daten:
  - Angaben über verschiedenste Fakten oder Zusammenhänge
  - Objektiv wahrnehmbar und potentiell verwertbar
- Informationen:
  - Erkennbare Semantik
  - Beseitigung einer Ungewissheit
- Finanzwelt:
  - Einzelner Aktienkurswert: *Datum*
  - Kurswert einer Firma zu mehreren Zeitpunkten: *Information* bei entsprechender Fachkenntnis

Informationsverarbeitung

- Daten $\neq$ Informationen
- Daten:
  - Angaben über verschiedenste Fakten oder Zusammenhänge
  - Objektiv wahrnehmbar und potentiell verwertbar
- Informationen:
  - Erkennbare Semantik
  - Beseitigung einer Ungewissheit
- Finanzwelt:
  - Einzelner Aktienkurswert: *Datum*
  - Kurswert einer Firma zu mehreren Zeitpunkten: *Information* bei entsprechender Fachkenntnis

Informationsverarbeitung

- Daten $\neq$ Informationen
- Daten:
  - Angaben über verschiedenste Fakten oder Zusammenhänge
  - Objektiv wahrnehmbar und potentiell verwertbar
- Informationen:
  - Erkennbare Semantik
  - Beseitigung einer Ungewissheit
- Finanzwelt:
  - Einzelner Aktienkurswert: *Datum*
  - Kurswert einer Firma zu mehreren Zeitpunkten: *Information* bei entsprechender Fachkenntnis

Textanalyse im Börsenumfeld

- Nachrichten aus dem Umfeld von Wirtschaft und Börse
  - Meldungen zu bestimmten Firmen oder speziellen Wirtschaftsbereichen
  - Neuigkeiten, die die gesamte Wirtschaft und die Finanzstabilität im Land betreffen
- Einfache und schnelle Verfügbarkeit über das Internet
- Menge der digital zur Verfügung stehenden Nachrichten macht es Aktionären/Börsenmaklern nahezu unmöglich, diese manuell zu analysieren und damit zeitnah und angemessen reagieren zu können
- Auswertung mit Hilfe von Computersystemen gewünscht

Textanalyse im Börsenumfeld

- Nachrichten aus dem Umfeld von Wirtschaft und Börse
  - Meldungen zu bestimmten Firmen oder speziellen Wirtschaftsbereichen
  - Neuigkeiten, die die gesamte Wirtschaft und die Finanzstabilität im Land betreffen
- Einfache und schnelle Verfügbarkeit über das Internet
- Menge der digital zur Verfügung stehenden Nachrichten macht es Aktionären/Börsenmaklern nahezu unmöglich, diese manuell zu analysieren und damit zeitnah und angemessen reagieren zu können
- Auswertung mit Hilfe von Computersystemen gewünscht

Textanalyse im Börsenumfeld

- Nachrichten aus dem Umfeld von Wirtschaft und Börse
  - Meldungen zu bestimmten Firmen oder speziellen Wirtschaftsbereichen
  - Neuigkeiten, die die gesamte Wirtschaft und die Finanzstabilität im Land betreffen
- Einfache und schnelle Verfügbarkeit über das Internet
- Menge der digital zur Verfügung stehenden Nachrichten macht es Aktionären/Börsenmaklern nahezu unmöglich, diese manuell zu analysieren und damit zeitnah und angemessen reagieren zu können
- Auswertung mit Hilfe von Computersystemen gewünscht

Textanalyse im Börsenumfeld

- Nachrichten aus dem Umfeld von Wirtschaft und Börse
  - Meldungen zu bestimmten Firmen oder speziellen Wirtschaftsbereichen
  - Neuigkeiten, die die gesamte Wirtschaft und die Finanzstabilität im Land betreffen
- Einfache und schnelle Verfügbarkeit über das Internet
- Menge der digital zur Verfügung stehenden Nachrichten macht es Aktionären/Börsenmaklern nahezu unmöglich, diese manuell zu analysieren und damit zeitnah und angemessen reagieren zu können
- Auswertung mit Hilfe von Computersystemen gewünscht

Bestehende Software und Erweiterungswünsche

- In Studienarbeit „Nachrichtenklassikation unter Nutzung regulärer Ausdrücke“ Prototyp entstanden
- Umfangreiche Vorverarbeitung von deutschsprachigen Börsennachrichten und Naive-Bayes-Klassifikation → prozentuelle Vorhersage *UP*/*DOWN*
- Nutzung einer Java-Statistikbibliothek, einzelne Bearbeitungsschritte mit `bash`-Skripten implementiert
- Klassifizierungsmöglichkeit neuer unbekannter Texte mit bereits erlernten Klassifikatoren und benutzerfreundliche grafische Oberfläche gewünscht
- Integration dieser Software in bestehendes Informationssystem WEBIS gewünscht

Bestehende Software und Erweiterungswünsche

- In Studienarbeit „Nachrichtenklassikation unter Nutzung regulärer Ausdrücke“ Prototyp entstanden
- Umfangreiche Vorverarbeitung von deutschsprachigen Börsennachrichten und Naive-Bayes-Klassifikation $\longrightarrow$ prozentuelle Vorhersage *UP*/*DOWN*
- Nutzung einer Java-Statistikbibliothek, einzelne Bearbeitungsschritte mit `bash`-Skripten implementiert
- Klassifizierungsmöglichkeit neuer unbekannter Texte mit bereits erlernten Klassifikatoren und benutzerfreundliche grafische Oberfläche gewünscht
- Integration dieser Software in bestehendes Informationssystem WEBIS gewünscht

Bestehende Software und Erweiterungswünsche

- In Studienarbeit „Nachrichtenklassikation unter Nutzung regulärer Ausdrücke“ Prototyp entstanden
- Umfangreiche Vorverarbeitung von deutschsprachigen Börsennachrichten und Naive-Bayes-Klassifikation $\longrightarrow$ prozentuelle Vorhersage *UP*/*DOWN*
- Nutzung einer Java-Statistikbibliothek, einzelne Bearbeitungsschritte mit `bash`-Skripten implementiert
- Klassifizierungsmöglichkeit neuer unbekannter Texte mit bereits erlernten Klassifikatoren und benutzerfreundliche grafische Oberfläche gewünscht
- Integration dieser Software in bestehendes Informationssystem WEBIS gewünscht

Bestehende Software und Erweiterungswünsche

- In Studienarbeit „Nachrichtenklassikation unter Nutzung regulärer Ausdrücke“ Prototyp entstanden
- Umfangreiche Vorverarbeitung von deutschsprachigen Börsennachrichten und Naive-Bayes-Klassifikation $\longrightarrow$ prozentuelle Vorhersage *UP*/*DOWN*
- Nutzung einer Java-Statistikbibliothek, einzelne Bearbeitungsschritte mit `bash`-Skripten implementiert
- Klassifizierungsmöglichkeit neuer unbekannter Texte mit bereits erlernten Klassifikatoren und benutzerfreundliche grafische Oberfläche gewünscht
- Integration dieser Software in bestehendes Informationssystem WEBIS gewünscht

Bestehende Software und Erweiterungswünsche

- In Studienarbeit „Nachrichtenklassikation unter Nutzung regulärer Ausdrücke“ Prototyp entstanden
- Umfangreiche Vorverarbeitung von deutschsprachigen Börsennachrichten und Naive-Bayes-Klassifikation $\longrightarrow$ prozentuelle Vorhersage *UP*/*DOWN*
- Nutzung einer Java-Statistikbibliothek, einzelne Bearbeitungsschritte mit `bash`-Skripten implementiert
- Klassifizierungsmöglichkeit neuer unbekannter Texte mit bereits erlernten Klassifikatoren und benutzerfreundliche grafische Oberfläche gewünscht
- Integration dieser Software in bestehendes Informationssystem WEBIS gewünscht

WEBIS-Historie

- 2001 Web-orientiertes Informationssystem entwickelt
  - Java-GUI mit Wrapper-Generator und XML-Repository
  - Datenspeicherung in XML
- Funktionalität wird mit Namen ausgedrückt
- Name = Projekt, Programm und Funktion
- Anfänglich keine Spezialisierung auf Themengebiet, später wegen umfangreicher Datenverfügbarkeit Börsenmarkt ausgewählt
- Diverse Architekturveränderungen:
  - Client/Server
  - Plug-in-Prinzip

WEBIS-Historie

- 2001 Web-orientiertes Informationssystem entwickelt
  - Java-GUI mit Wrapper-Generator und XML-Repository
  - Datenspeicherung in XML
- Funktionalität wird mit Namen ausgedrückt
- Name = Projekt, Programm und Funktion
- Anfänglich keine Spezialisierung auf Themengebiet, später wegen umfangreicher Datenverfügbarkeit Börsenmarkt ausgewählt
- Diverse Architekturveränderungen:
  - Client/Server
  - Plug-in-Prinzip

WEBIS-Historie

- 2001 Web-orientiertes Informationssystem entwickelt
  - Java-GUI mit Wrapper-Generator und XML-Repository
  - Datenspeicherung in XML
- Funktionalität wird mit Namen ausgedrückt
- Name = Projekt, Programm und Funktion
- Anfänglich keine Spezialisierung auf Themengebiet, später wegen umfangreicher Datenverfügbarkeit Börsenmarkt ausgewählt
- Diverse Architekturveränderungen:
  - Client/Server
  - Plug-in-Prinzip

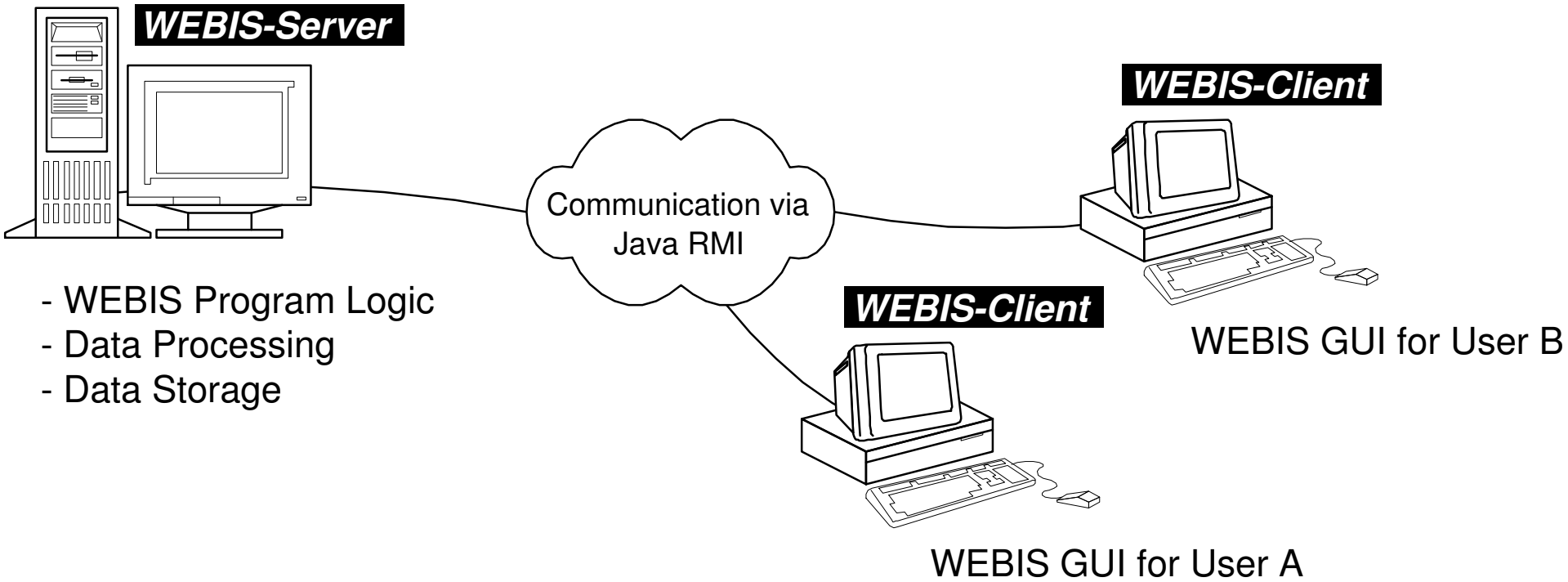
WEBIS-Historie

- 2001 Web-orientiertes Informationssystem entwickelt
  - Java-GUI mit Wrapper-Generator und XML-Repository
  - Datenspeicherung in XML
- Funktionalität wird mit Namen ausgedrückt
- Name = Projekt, Programm und Funktion
- Anfänglich keine Spezialisierung auf Themengebiet, später wegen umfangreicher Datenverfügbarkeit Börsenmarkt ausgewählt
- Diverse Architekturveränderungen:
  - Client/Server
  - Plug-in-Prinzip

WEBIS-Historie

- 2001 Web-orientiertes Informationssystem entwickelt
  - Java-GUI mit Wrapper-Generator und XML-Repository
  - Datenspeicherung in XML
- Funktionalität wird mit Namen ausgedrückt
- Name = Projekt, Programm und Funktion
- Anfänglich keine Spezialisierung auf Themengebiet, später wegen umfangreicher Datenverfügbarkeit Börsenmarkt ausgewählt
- Diverse Architekturveränderungen:
  - Client/Server
  - Plug-in-Prinzip

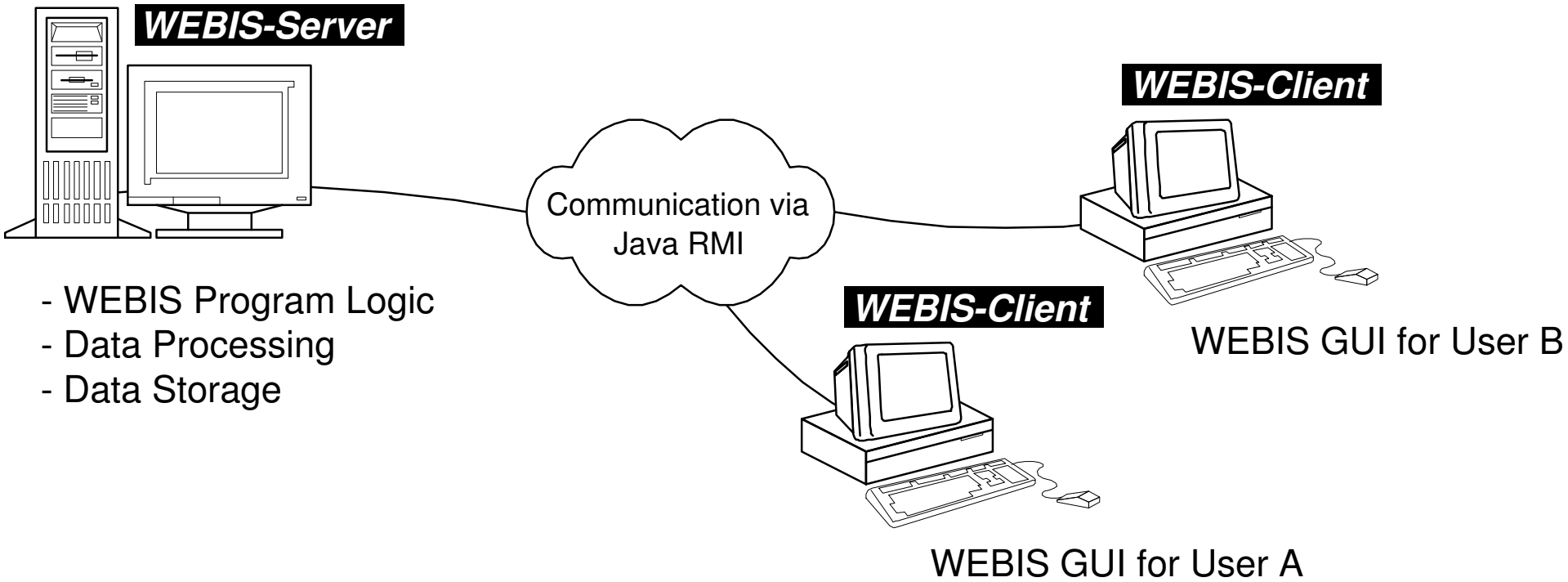
WEBIS-Client/Server-Architektur



Quelle: Diplomarbeit Frank Ahnert 2005

- Clients dienen fast ausschließlich als Benutzerschnittstelle
- Plug-in-Anbindung an Client und Server

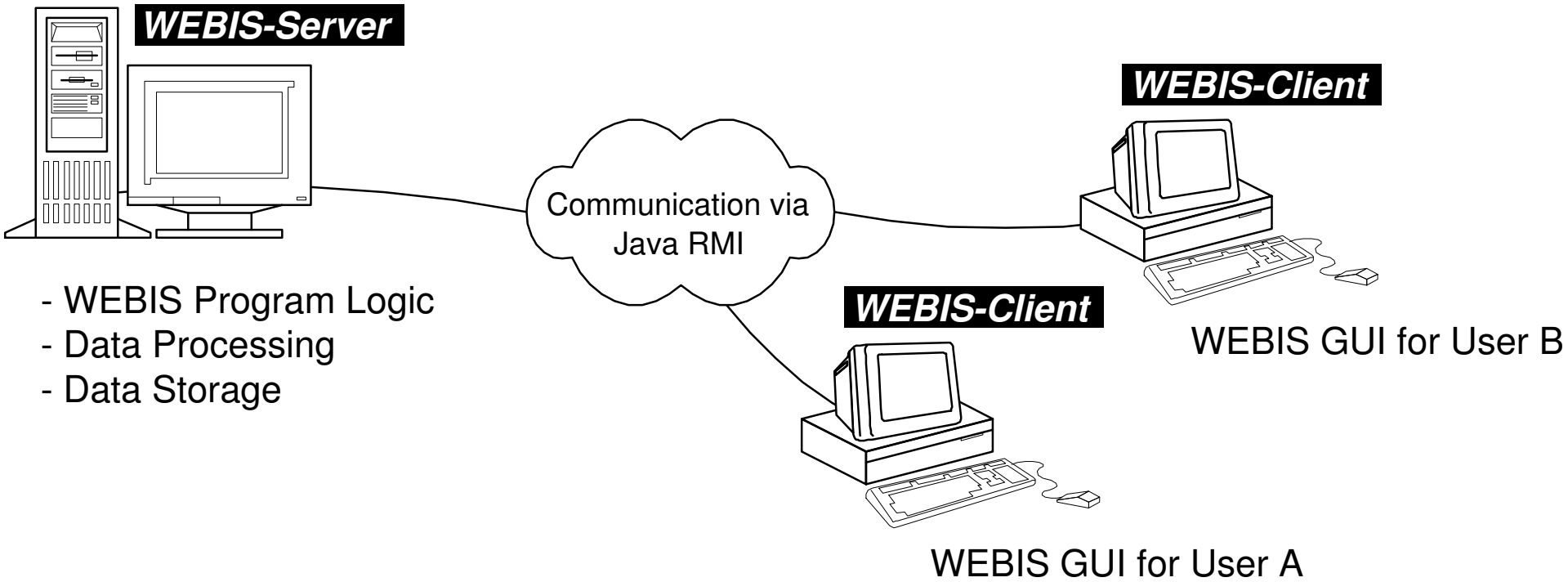
WEBIS-Client/Server-Architektur



Quelle: Diplomarbeit Frank Ahnert 2005

- Clients dienen fast ausschließlich als Benutzerschnittstelle
- Plug-in-Anbindung an Client und Server

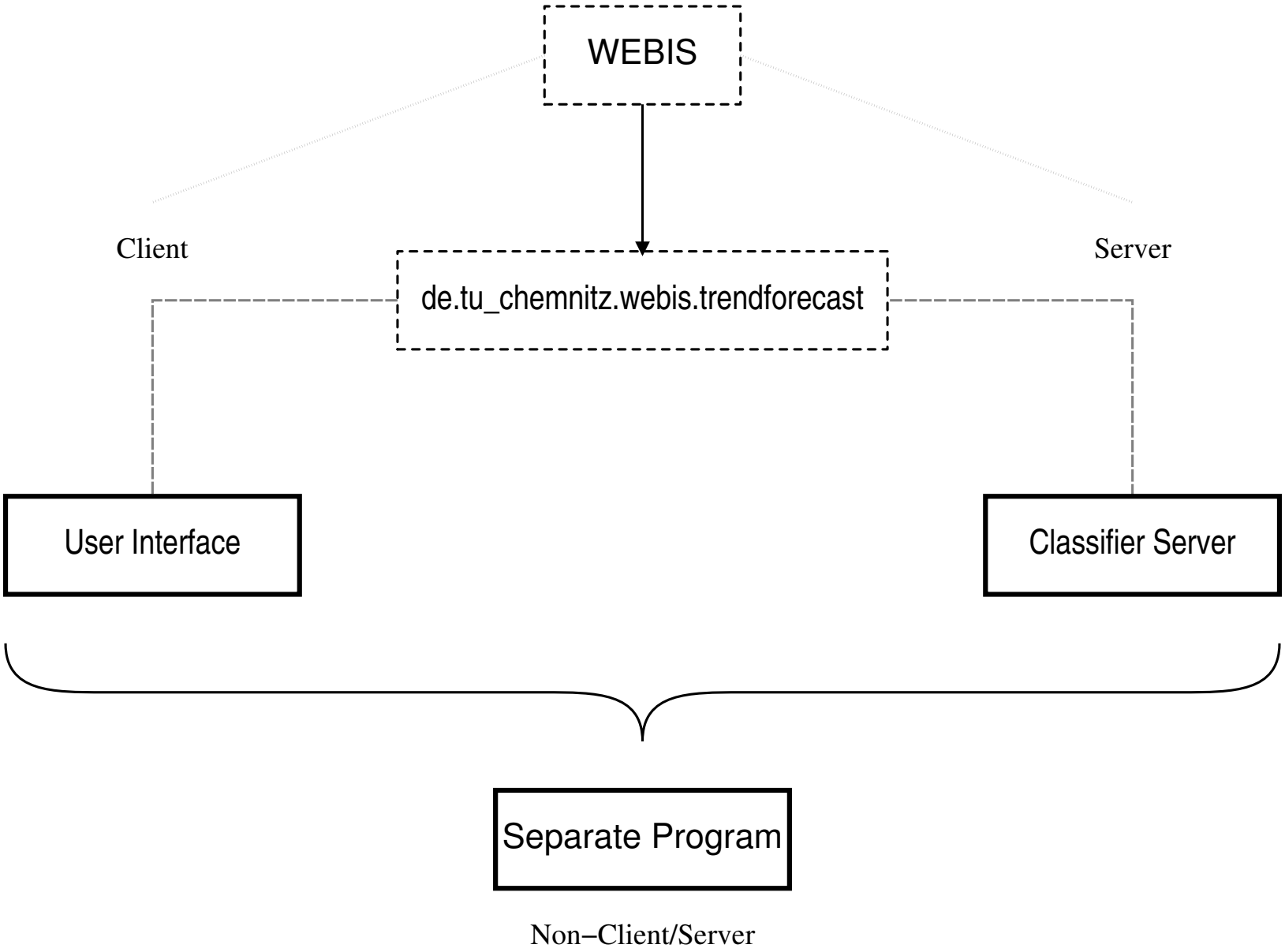
WEBIS-Client/Server-Architektur



Quelle: Diplomarbeit Frank Ahnert 2005

- Clients dienen fast ausschließlich als Benutzerschnittstelle
- Plug-in-Anbindung an Client und Server

Übersicht der entwickelten Software



→ 2006-KW27/06-07-05_16-57_0000.txt

Börsengangs der russischen Ölgesellschaft OAO Rosneft unverbindlich Aktien gezeichnet. Der britische Ölkonzern "hat informellen geordert. Er könnte immer noch aussteigen", sagte am Mittwoch eine Person, die mit der Transaktion zu tun hat. DJG/flf/nas

05.07.06 16:57 *ALLGEMEINES*
MARKT/Euro-Stoxx-50 nahe 200-Tage-Linie
965814 *Dow Jones Euro STOXX* 50 *Index*
MARKT/Euro-Stoxx-50 nahe 200-Tage-Linie

Auf ein erneutes Unterschreiten der 200-Tage-Linie im Euro-Stoxx-50 stellen sich Marktteilnehmer ein. Sie liege bei 3.609 Punkten. Im DAX verlaufe die 38- Tage-Linie bei 5.579 und die 200-Tage-Linie bei 5.519 Punkten. Euro-Stoxx-50 minus 1,6% auf 3.610 Punkte. DJG/hru/raz

05.07.06 16:54 *UNTERNEHMEN INLAND*
MARKT/"Fat finger" in Siemens

→ 2006-KW27/06-07-05_16-57_0000.txt

Börsengangs der russischen Ölgesellschaft OAO Rosneft unverbindlich Aktien gezeichnet. Der britische Ölkonzern "hat informellen geordert. Er könnte immer noch aussteigen", sagte am Mittwoch eine Person, die mit der Transaktion zu tun hat. DJG/flf/nas

05.07.06 16:57 *ALLGEMEINES*
MARKT/Euro-Stoxx-50 nahe 200-Tage-Linie

965814 *Dow Jones Euro STOXX* 50 *Index*
MARKT/Euro-Stoxx-50 nahe 200-Tage-Linie

Auf ein erneutes Unterschreiten der 200-Tage-Linie im Euro-Stoxx-50 stellen sich Marktteilnehmer ein. Sie liege bei 3.609 Punkten. Im DAX verlaufe die 38- Tage-Linie bei 5.579 und die 200-Tage-Linie bei 5.519 Punkten. Euro-Stoxx-50 minus 1,6% auf 3.610 Punkte. DJG/hru/raz

---

05.07.06 16:54 *UNTERNEHMEN INLAND*
MARKT/"Fat finger" in Siemens

→ 2006-KW27/06-07-05_16-54_0000.txt

723610 *Siemens NA*
MARKT/"Fat finger" in Siemens

Ein "fat finger" hat Siemens nach unten gedrückt,
seitdem erholt sich der Kurs nach
Vola-Interruption. Derzeit steht der Kurs 2,1%
im Minus bei 67,13 *EUR nach einem Tief von* 66,40
EUR. DJG/hru/raz

05.07.06 16:53 *ALLGEMEINES*
MARKT/DAX fällt weiter - JPM nimmt Arbei
846900 *DAX*
MARKT/DAX fällt weiter - JPM nimmt Arbeitsmarktprognose hoch

Der DAX fällt auf neue Tagestiefs und steht nur noch
knapp über 5.600 *Punkten. J.P. Morgan habe nach*
dem ADP-Bericht nun die Prognose für den
Arbeitsmarktbericht am Freitag nach oben
genommen, das Haus erwarte nun 225.000 *neue*
Stellen außerhalb der Landwirtschaft nach

→ 2006-KW27/06-07-05_16-54_0000.txt

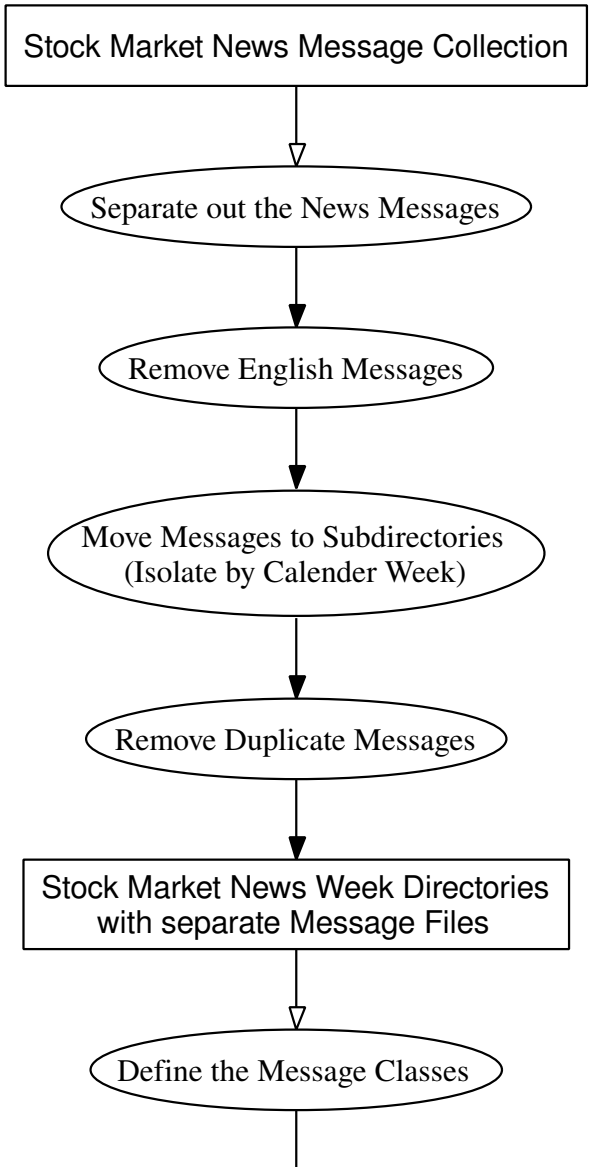
723610 *Siemens NA*
MARKT/"Fat finger" in Siemens

Ein "fat finger" hat Siemens nach unten gedrückt,
seitdem erholt sich der Kurs nach
Vola-Interruption. Derzeit steht der Kurs 2,1%
im Minus bei 67,13 *EUR nach einem Tief von* 66,40
EUR. DJG/hru/raz

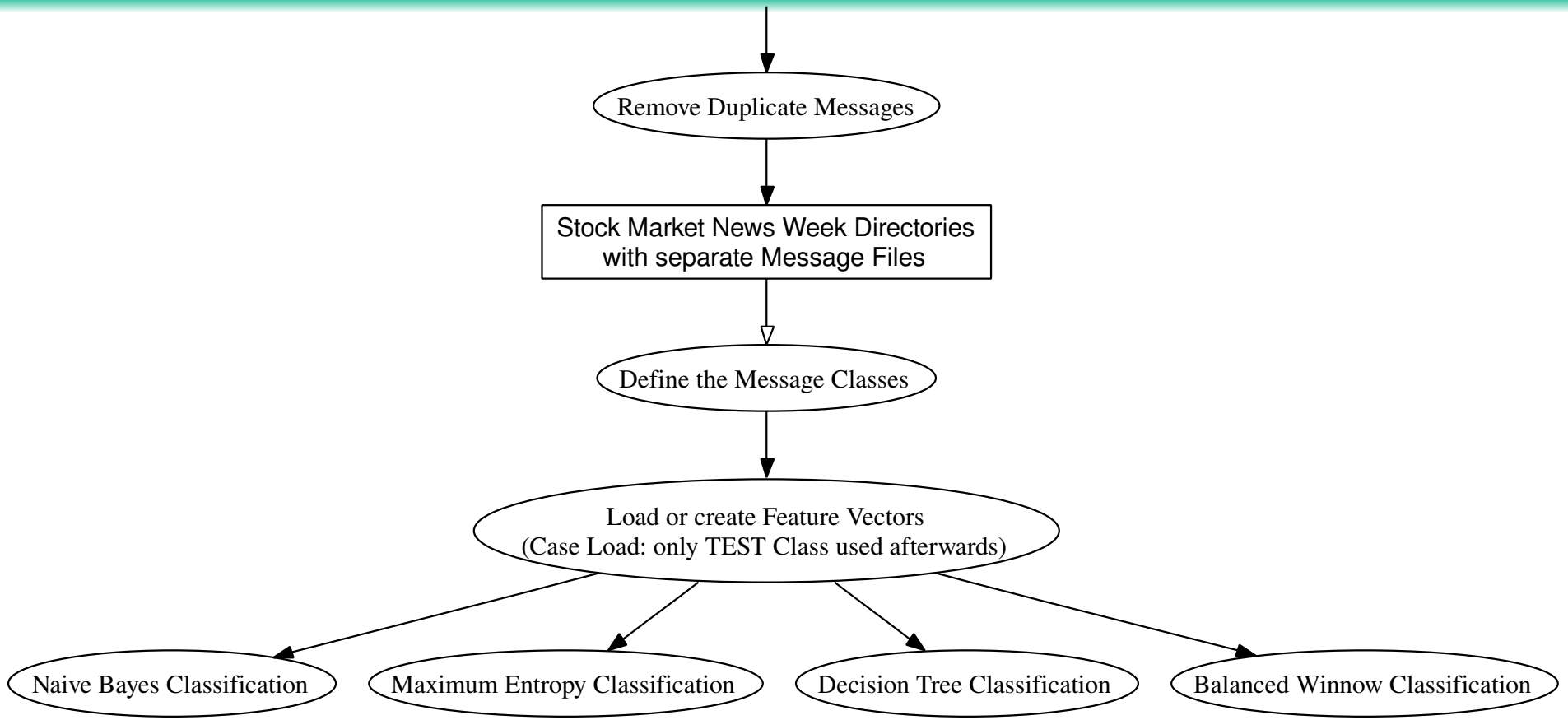
05.07.06 16:53 *ALLGEMEINES*
MARKT/DAX fällt weiter - JPM nimmt Arbei
846900 *DAX*
MARKT/DAX fällt weiter - JPM nimmt Arbeitsmarktprognose hoch

Der DAX fällt auf neue Tagestiefs und steht nur noch
knapp über 5.600 *Punkten. J.P. Morgan habe nach*
dem ADP-Bericht nun die Prognose für den
Arbeitsmarktbericht am Freitag nach oben
genommen, das Haus erwarte nun 225.000 *neue*
Stellen außerhalb der Landwirtschaft nach

Prinzipieller Bearbeitungsablauf I



Prinzipieller Bearbeitungsablauf II



Bemerkung

Klassifikation mit 4 von der Charakteristik her gesehen recht unterschiedlichen Classifiern möglich

Naive Bayes

- Abgeleitet aus Satz von Bayes
- Thomas Bayes' Satz zur Rechnung mit bedingten Wahrscheinlichkeiten: $P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$
- Gilt für endlich viele Ereignisse A_i:

$$P(A_i|B) = \frac{P(B|A_i) \cdot P(A_i)}{P(B)} = \frac{P(B|A_i) \cdot P(A_i)}{\sum_{j=1}^{N} P(B|A_j) \cdot P(A_j)}$$

- Bayes-Klassifikator ordnet zu klassifizierendem Objekt wahrscheinlichste Klasse zu (für die *UP/DOWN*-Klassifizierung gilt $C := \{0, 1\}$):

$$b : \mathbb{R}^f \rightarrow C$$

- Wahrscheinlichkeitsdichte aller i, $i \in C$ durch p_i gegeben $\rightarrow$ Klassenzugehörigkeit definiert durch

$$b := \texttt{argmax}_i \; p_i(x)$$

Naive Bayes

- Abgeleitet aus Satz von Bayes
- Thomas Bayes' Satz zur Rechnung mit bedingten Wahrscheinlichkeiten: $P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$
- Gilt für endlich viele Ereignisse A_i:

$$P(A_i|B) = \frac{P(B|A_i) \cdot P(A_i)}{P(B)} = \frac{P(B|A_i) \cdot P(A_i)}{\sum_{j=1}^{N} P(B|A_j) \cdot P(A_j)}$$

- Bayes-Klassifikator ordnet zu klassifizierendem Objekt wahrscheinlichste Klasse zu (für die *UP/DOWN*-Klassifizierung gilt $C := \{0, 1\}$):

$$b : \mathbb{R}^f \rightarrow C$$

- Wahrscheinlichkeitsdichte aller i, $i \in C$ durch p_i gegeben $\rightarrow$ Klassenzugehörigkeit definiert durch

$$b := \texttt{argmax}_i \; p_i(x)$$

Naive Bayes

- Abgeleitet aus Satz von Bayes
- Thomas Bayes' Satz zur Rechnung mit bedingten Wahrscheinlichkeiten: $P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$
- Gilt für endlich viele Ereignisse A_i:

$$P(A_i|B) = \frac{P(B|A_i) \cdot P(A_i)}{P(B)} = \frac{P(B|A_i) \cdot P(A_i)}{\sum_{j=1}^{N} P(B|A_j) \cdot P(A_j)}$$

- Bayes-Klassifikator ordnet zu klassifizierendem Objekt wahrscheinlichste Klasse zu (für die *UP/DOWN*-Klassifizierung gilt $C := \{0, 1\}$):

$$b : \mathbb{R}^f \rightarrow C$$

- Wahrscheinlichkeitsdichte aller i, $i \in C$ durch p_i gegeben $\rightarrow$ Klassenzugehörigkeit definiert durch

$$b := \texttt{argmax}_i \; p_i(x)$$

Naive Bayes

- Abgeleitet aus Satz von Bayes
- Thomas Bayes' Satz zur Rechnung mit bedingten Wahrscheinlichkeiten: $P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$
- Gilt für endlich viele Ereignisse A_i:

$$P(A_i|B) = \frac{P(B|A_i) \cdot P(A_i)}{P(B)} = \frac{P(B|A_i) \cdot P(A_i)}{\sum_{j=1}^{N} P(B|A_j) \cdot P(A_j)}$$

- Bayes-Klassifikator ordnet zu klassifizierendem Objekt wahrscheinlichste Klasse zu (für die *UP/DOWN*-Klassifizierung gilt $C := \{0, 1\}$):

$$b : \mathbb{R}^f \rightarrow C$$

- Wahrscheinlichkeitsdichte aller i, $i \in C$ durch p_i gegeben $\rightarrow$ Klassenzugehörigkeit definiert durch

$$b := \texttt{argmax}_i \; p_i(x)$$

Naive Bayes

- Abgeleitet aus Satz von Bayes
- Thomas Bayes' Satz zur Rechnung mit bedingten Wahrscheinlichkeiten: $P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$
- Gilt für endlich viele Ereignisse A_i:

$$P(A_i|B) = \frac{P(B|A_i) \cdot P(A_i)}{P(B)} = \frac{P(B|A_i) \cdot P(A_i)}{\sum_{j=1}^{N} P(B|A_j) \cdot P(A_j)}$$

- Bayes-Klassifikator ordnet zu klassifizierendem Objekt wahrscheinlichste Klasse zu (für die *UP/DOWN*-Klassifizierung gilt $C := \{0, 1\}$):

$$b : \mathbb{R}^f \rightarrow C$$

- Wahrscheinlichkeitsdichte aller i, $i \in C$ durch p_i gegeben $\rightarrow$ Klassenzugehörigkeit definiert durch

$$b := \texttt{argmax}_i \; p_i(x)$$

Maximum Entropy

- Verwendung zur optimalen Extraktion von Information aus verrauschten Signalen
- Prinzip der maximalen Entropie von Edwin Thompson Jaynes 1957 in die statistische Mechanik eingeführt
- Weiterentwickelte Methode findet Anwendung in Spektralanalyse und in der digitalen Bildverarbeitung
- Regularisierungs-/Glättungsmethode für die numerische Lösung von Problemen mit verrauschten Daten

Maximum Entropy

- Verwendung zur optimalen Extraktion von Information aus verrauschten Signalen
- Prinzip der maximalen Entropie von Edwin Thompson Jaynes 1957 in die statistische Mechanik eingeführt
- Weiterentwickelte Methode findet Anwendung in Spektralanalyse und in der digitalen Bildverarbeitung
- Regularisierungs-/Glättungsmethode für die numerische Lösung von Problemen mit verrauschten Daten

Maximum Entropy

- Verwendung zur optimalen Extraktion von Information aus verrauschten Signalen
- Prinzip der maximalen Entropie von Edwin Thompson Jaynes 1957 in die statistische Mechanik eingeführt
- Weiterentwickelte Methode findet Anwendung in Spektralanalyse und in der digitalen Bildverarbeitung
- Regularisierungs-/Glättungsmethode für die numerische Lösung von Problemen mit verrauschten Daten

Maximum Entropy

- Verwendung zur optimalen Extraktion von Information aus verrauschten Signalen
- Prinzip der maximalen Entropie von Edwin Thompson Jaynes 1957 in die statistische Mechanik eingeführt
- Weiterentwickelte Methode findet Anwendung in Spektralanalyse und in der digitalen Bildverarbeitung
- Regularisierungs-/Glättungsmethode für die numerische Lösung von Problemen mit verrauschten Daten

Balanced Winnow

- Von Nick Littlestone 1987 entwickeltes statistisches Verfahren
- Kalkuliert Wertigkeit jeder Klasse statt Wahrscheinlichkeiten direkt zu berechnen
- *Balanced-Winnow-Verfahren* verwendet zwei Gewichte für jedes Feature, ω^+ und ω^-.
- $$\Omega = \sum_{j=1}^{n} (\omega_j^+ - \omega_j^-) s_j > \theta, \text{ initial: } \omega^+ = \frac{2\theta}{d} \text{ und } \omega^- = \frac{\theta}{d}$$

  (d: durchschnittliche Anzahl Features pro Dokument, s_j: Feature j einer Dokumentinstanz $s = (s_1, s_2, \ldots, s_n)$)
- Beförderungsfaktor α und Degradierungsfaktor β werden jeweils gleichzeitig angewandt, um Klassifikator zu trainieren

Balanced Winnow

- Von Nick Littlestone 1987 entwickeltes statistisches Verfahren
- Kalkuliert Wertigkeit jeder Klasse statt Wahrscheinlichkeiten direkt zu berechnen
- *Balanced-Winnow-Verfahren* verwendet zwei Gewichte für jedes Feature, ω^+ und ω^-.
- $$\Omega = \sum_{j=1}^{n} (\omega_j^+ - \omega_j^-) s_j > \theta, \text{ initial: } \omega^+ = \frac{2\theta}{d} \text{ und } \omega^- = \frac{\theta}{d}$$

  (d: durchschnittliche Anzahl Features pro Dokument, s_j: Feature j einer Dokumentinstanz $s = (s_1, s_2, \ldots, s_n)$)
- Beförderungsfaktor α und Degradierungsfaktor β werden jeweils gleichzeitig angewandt, um Klassifikator zu trainieren

Balanced Winnow

- Von Nick Littlestone 1987 entwickeltes statistisches Verfahren
- Kalkuliert Wertigkeit jeder Klasse statt Wahrscheinlichkeiten direkt zu berechnen
- *Balanced-Winnow-Verfahren* verwendet zwei Gewichte für jedes Feature, ω^+ und ω^-.
- $$\Omega = \sum_{j=1}^{n} (\omega_j^+ - \omega_j^-) s_j > \theta, \text{ initial: } \omega^+ = \frac{2\theta}{d} \text{ und } \omega^- = \frac{\theta}{d}$$

  (d: durchschnittliche Anzahl Features pro Dokument, s_j: Feature j einer Dokumentinstanz $s = (s_1, s_2, \ldots, s_n)$)
- Beförderungsfaktor α und Degradierungsfaktor β werden jeweils gleichzeitig angewandt, um Klassifikator zu trainieren

Balanced Winnow

- Von Nick Littlestone 1987 entwickeltes statistisches Verfahren
- Kalkuliert Wertigkeit jeder Klasse statt Wahrscheinlichkeiten direkt zu berechnen
- *Balanced-Winnow-Verfahren* verwendet zwei Gewichte für jedes Feature, ω^+ und ω^-.
- $$\Omega = \sum_{j=1}^{n} (\omega_j^+ - \omega_j^-) s_j > \theta, \text{ initial: } \omega^+ = \frac{2\theta}{d} \text{ und } \omega^- = \frac{\theta}{d}$$

  (d: durchschnittliche Anzahl Features pro Dokument, s_j: Feature j einer Dokumentinstanz $s = (s_1, s_2, \ldots, s_n)$)
- Beförderungsfaktor α und Degradierungsfaktor β werden jeweils gleichzeitig angewandt, um Klassifikator zu trainieren

Balanced Winnow

- Von Nick Littlestone 1987 entwickeltes statistisches Verfahren
- Kalkuliert Wertigkeit jeder Klasse statt Wahrscheinlichkeiten direkt zu berechnen
- *Balanced-Winnow-Verfahren* verwendet zwei Gewichte für jedes Feature, ω^+ und ω^-.
- $$\Omega = \sum_{j=1}^{n} (\omega_j^+ - \omega_j^-) s_j > \theta, \text{ initial: } \omega^+ = \frac{2\theta}{d} \text{ und } \omega^- = \frac{\theta}{d}$$

  (d: durchschnittliche Anzahl Features pro Dokument, s_j: Feature j einer Dokumentinstanz $s = (s_1, s_2, \ldots, s_n)$)
- Beförderungsfaktor α und Degradierungsfaktor β werden jeweils gleichzeitig angewandt, um Klassifikator zu trainieren

Decision Tree

- Anschauliche Darstellung aufeinanderfolgender hierarchischer Verzweigungen in Baumform
- Binärer Fall: Jede Verzweigung eine Ja/Nein-Entscheidung
- Generierung üblich im Top-Down-Prinzip, in jedem Schritt Suche nach bestem Split (Nutzung der Entropie oder des so genannten Gini-Index)
- Problem des Overfitting: Klassifikationsgüte nimmt mit zunehmender Baumkomplexität ab

Decision Tree

- Anschauliche Darstellung aufeinanderfolgender hierarchischer Verzweigungen in Baumform
- Binärer Fall: Jede Verzweigung eine Ja/Nein-Entscheidung
- Generierung üblich im Top-Down-Prinzip, in jedem Schritt Suche nach bestem Split (Nutzung der Entropie oder des so genannten Gini-Index)
- Problem des Overfitting: Klassifikationsgüte nimmt mit zunehmender Baumkomplexität ab

Decision Tree

- Anschauliche Darstellung aufeinanderfolgender hierarchischer Verzweigungen in Baumform
- Binärer Fall: Jede Verzweigung eine Ja/Nein-Entscheidung
- Generierung üblich im Top-Down-Prinzip, in jedem Schritt Suche nach bestem Split (Nutzung der Entropie oder des so genannten Gini-Index)
- Problem des Overfitting: Klassifikationsgüte nimmt mit zunehmender Baumkomplexität ab

Decision Tree

- Anschauliche Darstellung aufeinanderfolgender hierarchischer Verzweigungen in Baumform
- Binärer Fall: Jede Verzweigung eine Ja/Nein-Entscheidung
- Generierung üblich im Top-Down-Prinzip, in jedem Schritt Suche nach bestem Split (Nutzung der Entropie oder des so genannten Gini-Index)
- Problem des Overfitting: Klassifikationsgüte nimmt mit zunehmender Baumkomplexität ab

Softwaremerkmale

- Dient der Datenvorverarbeitung,
  - der Klassenerstellung
  - zum Erstellen der Vektorendatei
  - und zum Klassifizieren der Testklassen inklusiver neuer Texte
- Hauptfunktionalität (außer Visualisierung) auf Server bzw. in monolithischer Einzelanwendung
- Großer Teil gemeinsamer Code möglich (Einzelanwendung als RMI-Client/Server-Anwendung auszuführen um noch weniger Anpassungen vornehmen zu müssen nicht sinnvoll)

Softwaremerkmale

- Dient der Datenvorverarbeitung,
  - der Klassenerstellung
  - zum Erstellen der Vektorendatei
  - und zum Klassifizieren der Testklassen inklusiver neuer Texte
- Hauptfunktionalität (außer Visualisierung) auf Server bzw. in monolithischer Einzelanwendung
- Großer Teil gemeinsamer Code möglich (Einzelanwendung als RMI-Client/Server-Anwendung auszuführen um noch weniger Anpassungen vornehmen zu müssen nicht sinnvoll)

Softwaremerkmale

- Dient der Datenvorverarbeitung,
  - der Klassenerstellung
  - zum Erstellen der Vektorendatei
  - und zum Klassifizieren der Testklassen inklusiver neuer Texte
- Hauptfunktionalität (außer Visualisierung) auf Server bzw. in monolithischer Einzelanwendung
- Großer Teil gemeinsamer Code möglich (Einzelanwendung als RMI-Client/Server-Anwendung auszuführen um noch weniger Anpassungen vornehmen zu müssen nicht sinnvoll)

Softwaremerkmale

- Dient der Datenvorverarbeitung,
  - der Klassenerstellung
  - zum Erstellen der Vektorendatei
  - und zum Klassifizieren der Testklassen inklusiver neuer Texte
- Hauptfunktionalität (außer Visualisierung) auf Server bzw. in monolithischer Einzelanwendung
- Großer Teil gemeinsamer Code möglich (Einzelanwendung als RMI-Client/Server-Anwendung auszuführen um noch weniger Anpassungen vornehmen zu müssen nicht sinnvoll)

Softwaremerkmale

- Dient der Datenvorverarbeitung,
  - der Klassenerstellung
  - zum Erstellen der Vektorendatei
  - und zum Klassifizieren der Testklassen inklusiver neuer Texte
- Hauptfunktionalität (außer Visualisierung) auf Server bzw. in monolithischer Einzelanwendung
- Großer Teil gemeinsamer Code möglich (Einzelanwendung als RMI-Client/Server-Anwendung auszuführen um noch weniger Anpassungen vornehmen zu müssen nicht sinnvoll)

Softwaremerkmale

- Dient der Datenvorverarbeitung,
  - der Klassenerstellung
  - zum Erstellen der Vektorendatei
  - und zum Klassifizieren der Testklassen inklusiver neuer Texte
- Hauptfunktionalität (außer Visualisierung) auf Server bzw. in monolithischer Einzelanwendung
- Großer Teil gemeinsamer Code möglich (Einzelanwendung als RMI-Client/Server-Anwendung auszuführen um noch weniger Anpassungen vornehmen zu müssen nicht sinnvoll)

plugin_server.xml

```
<?xml version="1.0"?>

<plugin name="de.tu_chemnitz.webis.trendforecast" version="1.0">

 <bind-pluginSocket
   class="de.tu_chemnitz.webis.trendforecast.ClassifierServer"
   name="de.tu_chemnitz.webis.core.RemotePluginSocket"/>

 <classpath name="commons-collections-3.2.jar" />
 <classpath name="commons-configuration-1.2.jar" />
 <classpath name="commons-lang-2.1.jar" />
 <classpath name="commons-logging-1.1.jar" />
 <classpath name="mallet-deps.jar" />
 <classpath name="mallet.jar" />

</plugin>
```

plugin_client.xml

```
<?xml version="1.0"?>

<plugin name="de.tu_chemnitz.webis.trendforecast" version="1.0">

 <bind-pluginSocket class="" name="de.tu_chemnitz.webis.gui.
     components.navigation.NavigationPanelPluginSocket" />

 <navigationNode id="User Interface" controlledByParent="false"
     path="/Trend Forecast" doubleClickAction="de.tu_chemnitz.webis.
         gui.trendforecast.WebisMainWindow"/>

 <dependency name="de.tu_chemnitz.webis.gui.components.navigation"
     version="1.0" />
 <dependency name="de.tu_chemnitz.webis.trendforecast" version="1.0"
      type="remote" />

 <classpath name="jcommon-1.0.0.jar" />
 <classpath name="jfreechart-1.0.1.jar" />
 <classpath name="org-jdesktop-layout.jar" />

</plugin>
```

Feature Reduction

- Performance- und Arbeitsspeicherbedarf des Prototyps mit den vollständig erzeugten Feature Vectors recht hoch
- Klassifizierung einer neuen Trainingseinheit „Kalenderwoche“ auf dem CPU-Server `herkules.hrz` in etwa 1 Minute, wenn Feature Vectors im Speicher geladen waren
- Reduktion der Menge an Feature Vectors in das hier entwickelte System integriert
- Nutzt kalkulierten Informationsgehalt für Rangfolge, aus der die anteilmäßig am schlechtesten bewerteten Vektoren entfernt werden

Feature Reduction

- Performance- und Arbeitsspeicherbedarf des Prototyps mit den vollständig erzeugten Feature Vectors recht hoch
- Klassifizierung einer neuen Trainingseinheit „Kalenderwoche“ auf dem CPU-Server `herkules.hrz` in etwa 1 Minute, wenn Feature Vectors im Speicher geladen waren
- Reduktion der Menge an Feature Vectors in das hier entwickelte System integriert
- Nutzt kalkulierten Informationsgehalt für Rangfolge, aus der die anteilmäßig am schlechtesten bewerteten Vektoren entfernt werden

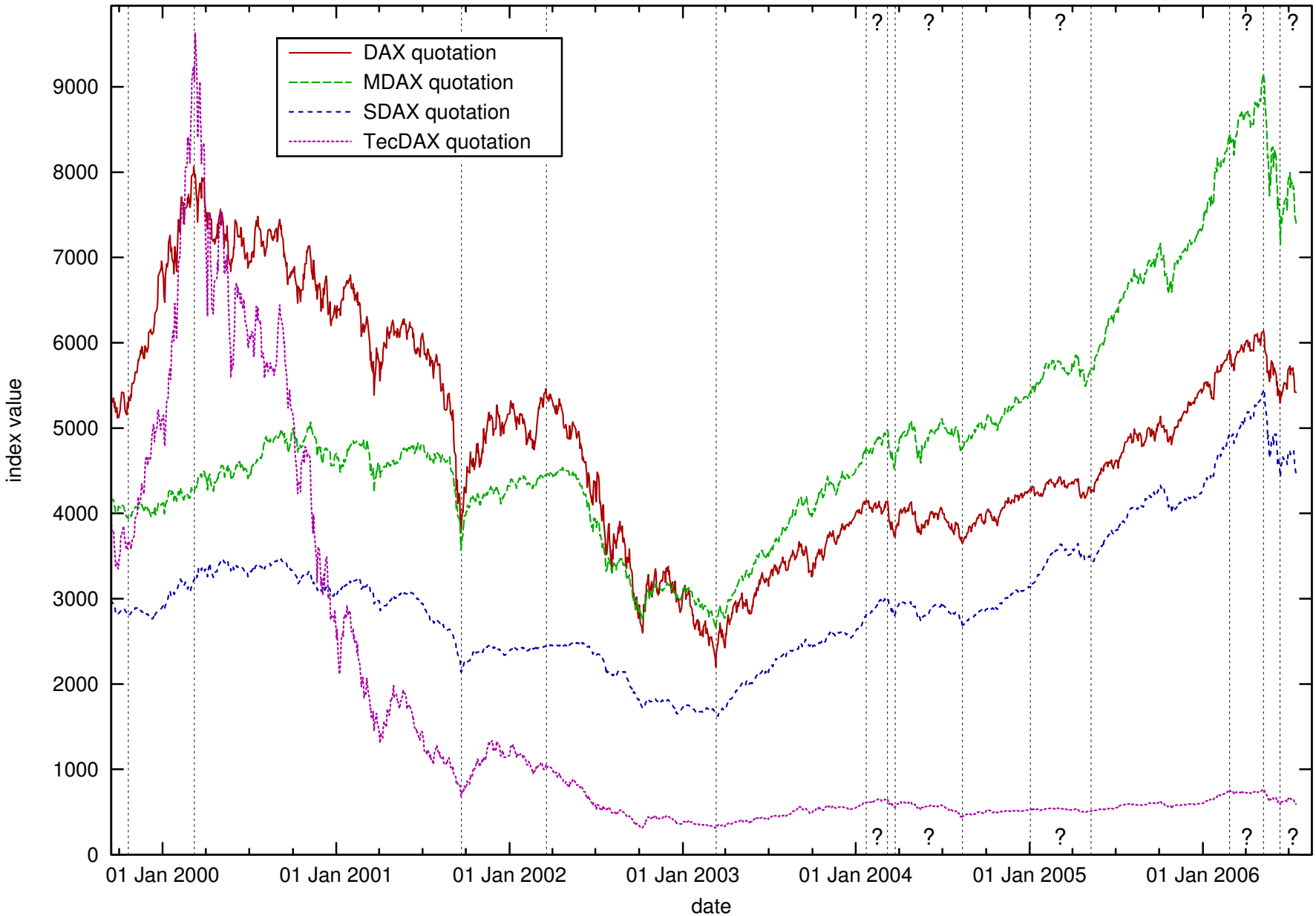
Feature Reduction

- Performance- und Arbeitsspeicherbedarf des Prototyps mit den vollständig erzeugten Feature Vectors recht hoch
- Klassifizierung einer neuen Trainingseinheit „Kalenderwoche“ auf dem CPU-Server `herkules.hrz` in etwa 1 Minute, wenn Feature Vectors im Speicher geladen waren
- Reduktion der Menge an Feature Vectors in das hier entwickelte System integriert
- Nutzt kalkulierten Informationsgehalt für Rangfolge, aus der die anteilmäßig am schlechtesten bewerteten Vektoren entfernt werden

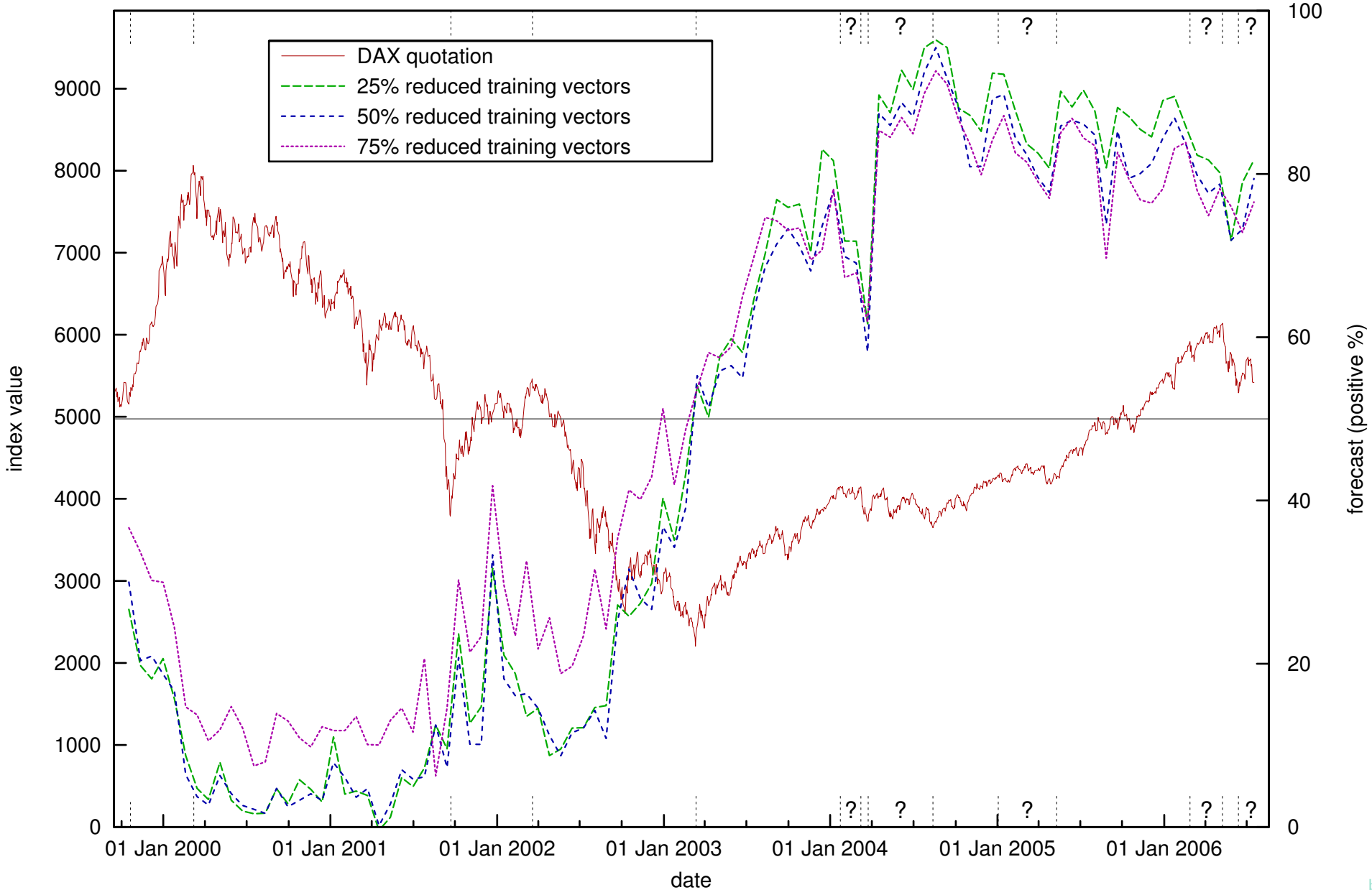
Feature Reduction

- Performance- und Arbeitsspeicherbedarf des Prototyps mit den vollständig erzeugten Feature Vectors recht hoch
- Klassifizierung einer neuen Trainingseinheit „Kalenderwoche“ auf dem CPU-Server `herkules.hrz` in etwa 1 Minute, wenn Feature Vectors im Speicher geladen waren
- Reduktion der Menge an Feature Vectors in das hier entwickelte System integriert
- Nutzt kalkulierten Informationsgehalt für Rangfolge, aus der die anteilmäßig am schlechtesten bewerteten Vektoren entfernt werden

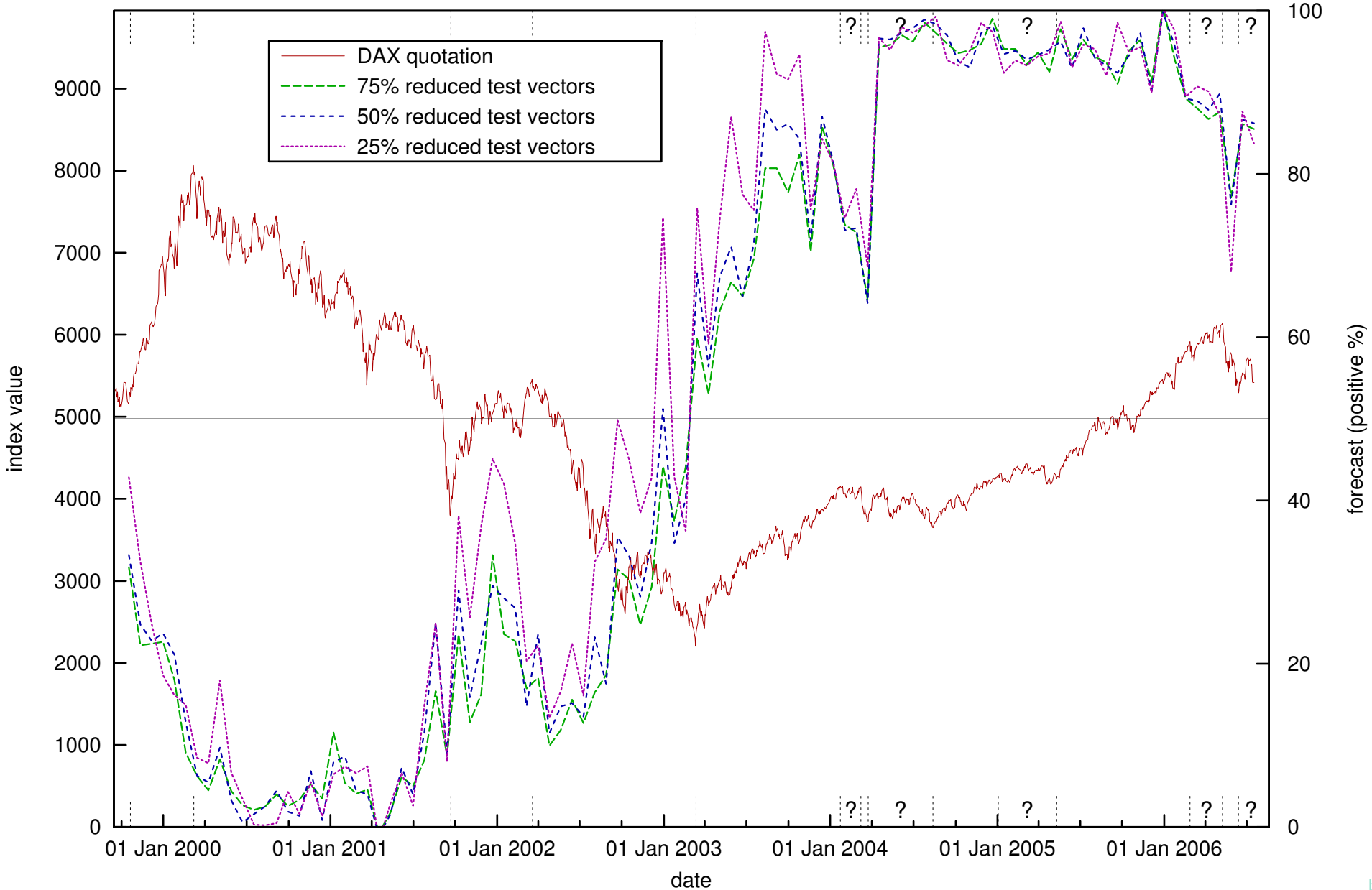
Für Klassifikatortraining definierte Phasen



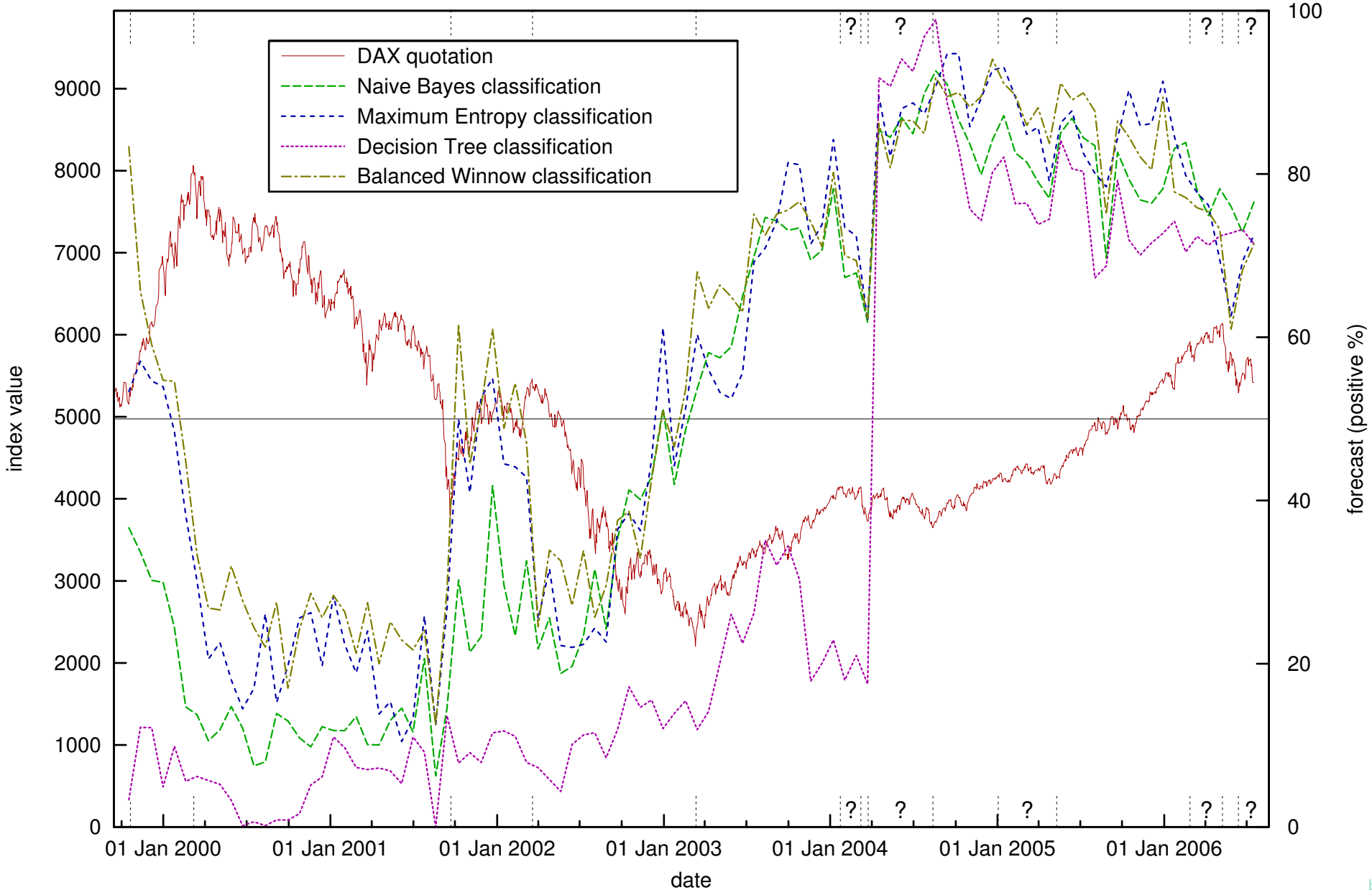
Naive-Bayes-Klassifikation, unreduzierte Testvektoren



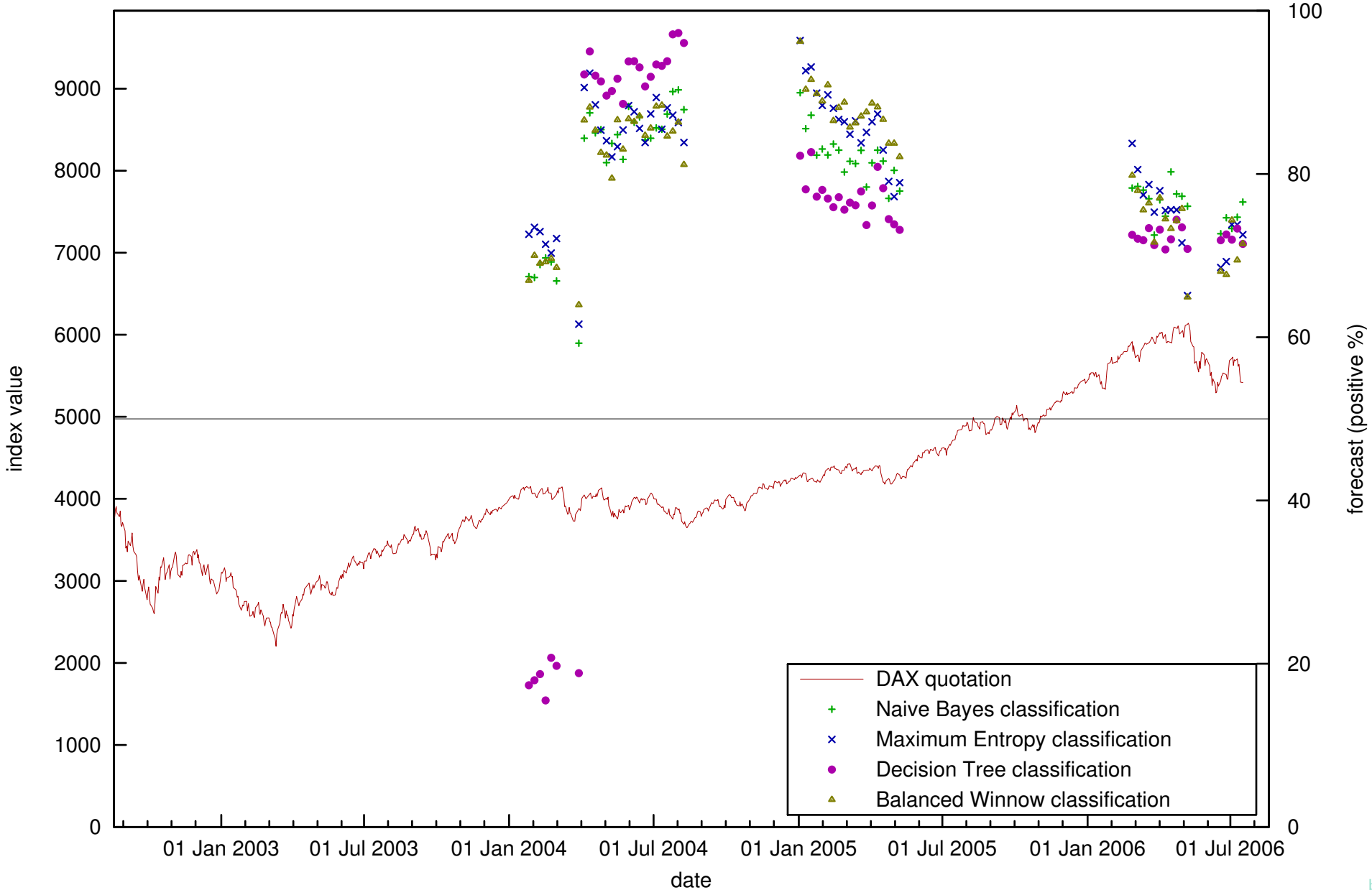
Naive-Bayes-Klassifikation, reduzierte Testvektoren



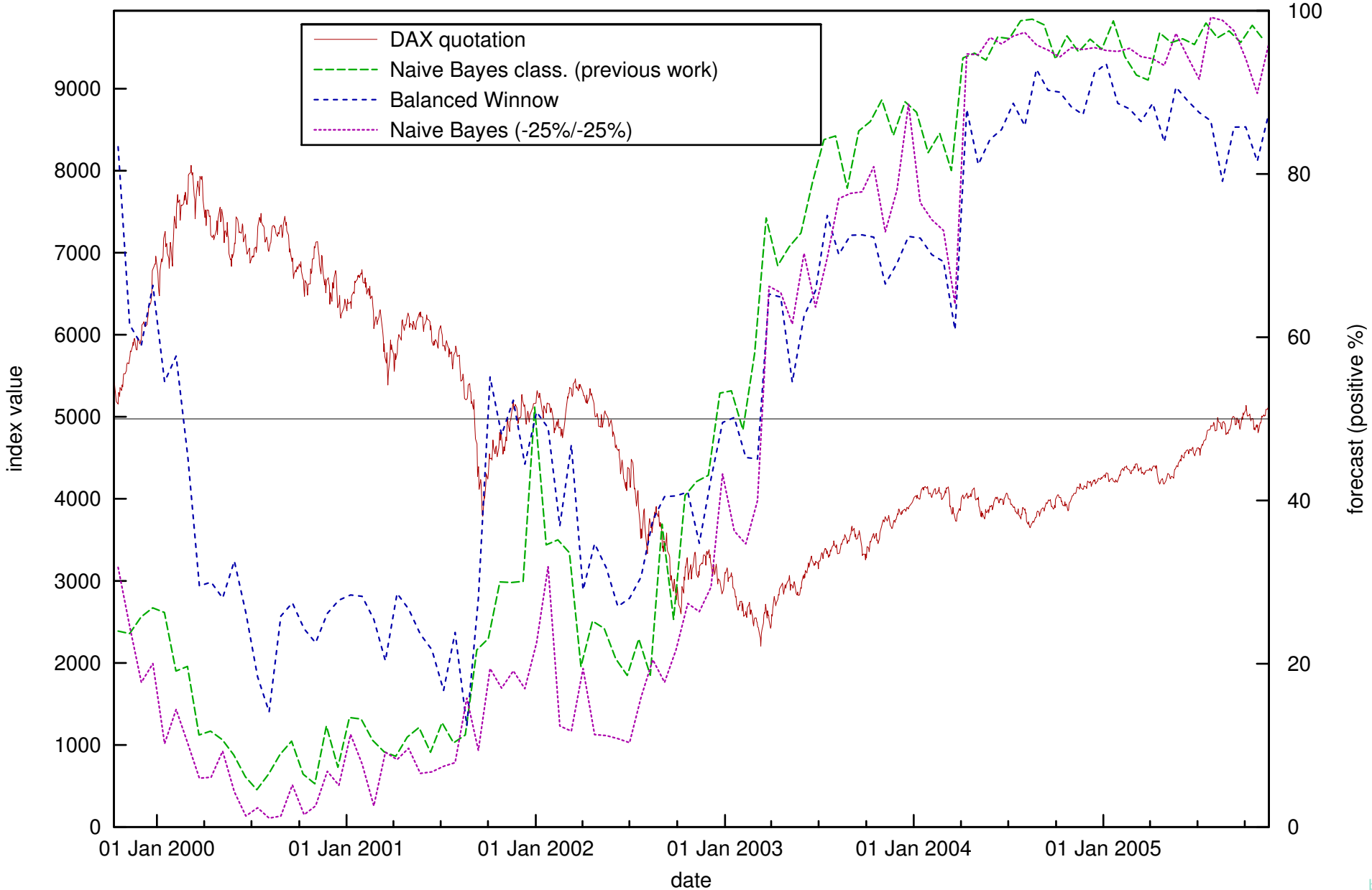
Klassifikationsergebnisse aller 4 verschiedenen Classifier



Testphasenanalyse aller 4 verschiedenen Classifier



Vergleich Klassifikation Studienarbeit mit neuen Klassifikat.



Vielen Dank für Ihr Interesse!