

From Human Language to Useful Information

Gilbane Boston
November 2007



BASIS
TECHNOLOGY

Steve Cohen

Founder and EVP

www.basistech.com

Matt Kodama

Product Management

www.endeca.com



ENDECA[®]

About Basis Technology

- Diversified firm specializing in language technology, text analytics, and software internationalization
- Twelve year track record delivering technology to and growing businesses overseas, including: Amazon.com, Ask.com, Google, Kenan Systems, L.L. Bean, Lycos, Ofoto, Overture, PeopleSoft, Progress Software, RightNow

Selected Basis Technology Customers

Google™

YAHOO!®



amazon.com.

Ask Jeeves™

楽天
ICHIWA

ENDECA®

fast

Verity™

CONVERA.

ORACLE®

EMC²
where information lives™

Tamino
XML Server

hp
invent

CISCO SYSTEMS
EMPOWERING THE
INTERNET GENERATION®

Information retrieval

- Finding the information you need

- ⇒ *Structured*

- Facts
 - Databases, Spreadsheets
 - Single field, row, join, etc. represents information of interest
 - Techniques for retrieval are well developed

- ⇒ *Unstructured*

- Text documents
 - Images
 - Audio, video, etc.
 - Also about facts, but these are harder to discern
 - Techniques for retrieval are less developed

Unstructured information retrieval

- Unstructured information retrieval begins and ends with human language



Human language is imprecise

- Choice of language
 - ⇒ *Esperanto, anyone?*
- Choice of writing quality
 - ⇒ *“Formal” news text*
 - ⇒ *“Informal” IM text*
 - CN U RD THIS?
- Choice of spelling
 - ⇒ *Mohammed*
 - ⇒ *Muhamed*
 - ⇒ *Mahmoud*



The screenshot shows the top portion of a news article on The New York Times website. The page header includes the newspaper's name, "The New York Times", and a navigation menu with categories: WORLD, U.S., N.Y. / REGION, BUSINESS, TECHNOLOGY, SCIENCE, and HEALTH. Below the navigation is a search bar with the text "Search Tech News & 8,000+ Products" and a "Go" button. To the right of the search bar is a "Browse Products" dropdown menu with the text "-- Select a Product". The main content area features the article title "Adding Math to List of Security Threats" in a large, bold font. Below the title, the author's name "By JOHN MARKOFF" and the publication date "Published: November 17, 2007" are displayed. The article's lead paragraph begins with "SAN FRANCISCO, Nov. 16 — One of the world's most prominent cryptographers issued a warning on Friday about a hypothetical incident in which a math error in a widely used computing chip pla the security of the global electronic commerce system at risk."

Technology to the rescue!

- Machine translation?
- Term translation
- Fuzzy matching
 - ⇒ *Large lists...*
 - ⇒ *Soundex, Phonix*
 - ⇒ *Edit Distance*
 - ⇒ *Machine learned models*
- Name finding



Machine Translation

Now I am alone.

O, what a rogue and peasant slave am I! (555)

Is it not monstrous that this player here,

But in a fiction, in a dream of passion,

Could force his soul so to his own conceit

That from her working all his visage wann'd,

Tears in his eyes, distraction in's aspect, (560)

A broken voice, and his whole function suiting

With forms to his conceit? and all for nothing!

For Hecuba!

Automatically translated text:

Now, I alone.

O, what a rogue and peasant slave me!

This is not monster, this player here,

However, in novels, in a dream of passion,

Could force his soul, and make our own conceited

This from her work, all his face wann'd,

Tears in his eyes, a part of the distraction,

A broken voice, and his overall functional adaptation

With forms to his conceited? And all why!

For hecuba!

English

|

Chinese

|

English

Term translation

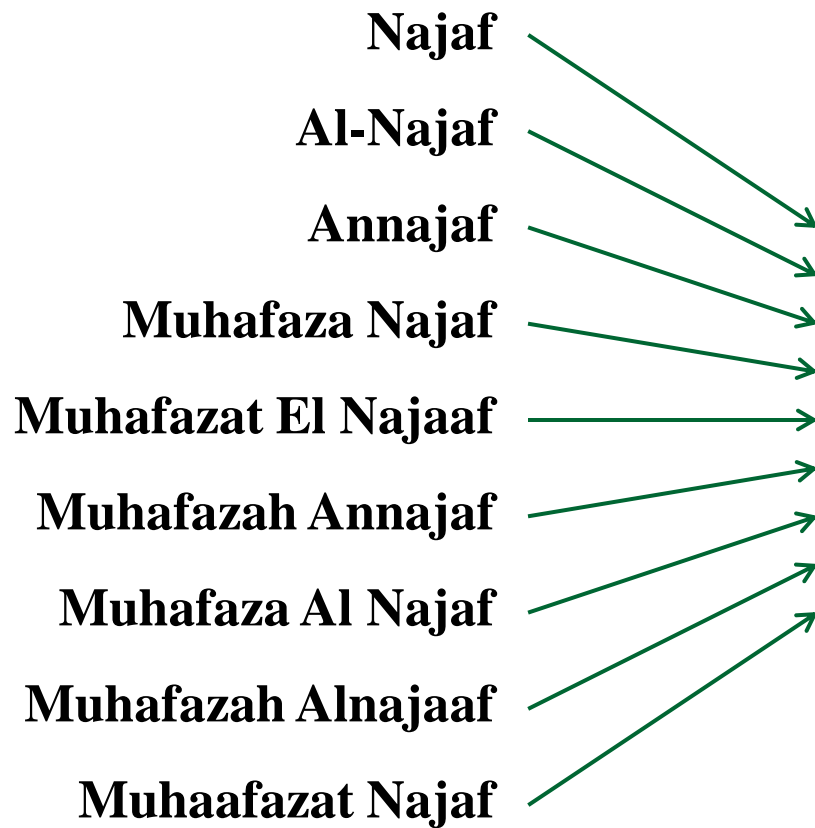
- “Taxi” in Chinese

- ⇒ 出租车

- ⇒ 计程车

- ⇒ 的士

Fuzzy Matching



محافظة النجف

Cross-Lingual Annotation

- Maintain original text
- Add information from target language
- Add linguistic meta-data
 - ⇒ *Entities (people, places, locations)*
 - ⇒ *Part of speech*
 - ⇒ *Word sense*

Rosette Text Analyzer

Dismiss

Results

Click on Arabic Words For Parse Data:

لرفع حالة الطوارئ المفروضة على Barwiz Musharraf al-Bakistani دعوتها للرئيس al-Wilayat al-Muttahidah جدد
بروز مشرف باكستاني الولايات المتحدة

بلادة منذ مطلع الشهر الجاري والعودة إلى النهج الدستوري.

Vocalized	Gloss	Part of Speech	Transliteration
أَلْوِلَايَات	(the) states/provinces ([fem.pl.])	DET NOUN FEM_PL	al-wilaayaat
أَلْوِلَايَات	(the) States ([fem.pl.])	DET NOUN FEM_PL	al-wilaayaat

Named Entities

LOCATION

الولايات المتحدة
al-Wilayat al-Muttahidah

باكستاني
al-Bakistani

PERSON

بروز مشرف
Barwiz Musharraf

Name Finding

- Find references to people, places, organizations
- Use this “semantic” knowledge about the text to enable retrieval or advanced analysis

The screenshot displays the Rosette Linguistics Platform interface. The main window shows a text document titled "en-text.txt" with the following metadata:

Language	English
MIME Type	text/plain
Encoding	UTF8
Length	1253

The text content is: "BBC News: Tuesday, 11 April 2006 Tokyo - The four-year-old daughter of Crown Prince Naruhito and his wife Princess Masako went to a beginning of term ceremony at Gakushuin".

Named Entities are listed in a sidebar:

- Person
- Organization
- Location
- Date
- Time
- Telephone
- Email Address
- URL
- Personal ID Number
- Credit Card Number
- Latitude/Longitude
- Money
- Percent
- Other

The main window also displays a table of named entities extracted from the text:

#	Type	Phrase
1	ORGANIZATION	BBC
2	TEMPORAL:DATE	Tuesday, 11 Apr...
3	LOCATION	Tokyo
4	PERSON	Naruhito
5	PERSON	Masako
6	ORGANIZATION	Gakushuin Kinde...
7	PERSON	Aiko
8	ORGANIZATION	Kyodo
9	LOCATION	Japan
1.	PERSON	Naruhito
1.	PERSON	Akishino
1.	PERSON	Junichiro Koizumi



Data Mining and Information Access

Matt Kodama
Product Management

Selected Endeca Customers

SallieMae®

ARROW
ARROW ELECTRONICS, INC.

OLD MUTUAL

Tech Data

Boston Scientific

COX
COMMUNICATIONS

Fidelity Investments

DOW CORNING

ABN·AMRO

Premier Farnell plc

BORDERS.

COSTCO
.COM

bhpbilliton

CREDIT SUISSE

HYATT®

CDW

JOHN DEERE

SAKS
INCORPORATED

IBM

Nike

RS

Marriott

Reed Elsevier

Walmart.com™

TEXAS INSTRUMENTS

HARRIS

Wyeth

AMERICAN EXPRESS

Boots

Life's Good LG

TESCO

CARMAX

otto group

circuits CITY

BARNES & NOBLE .COM
www.bn.com

John Hancock

INTERNATIONAL PAPER

TimeWarner

Agilent Technologies

LEHMAN BROTHERS

HENRY SCHEIN®

THE HOME DEPOT

Newell Rubbermaid

Office DEPOT

Walgreens

AutoZone.com™

HCR·ManorCare

SIGNET

K
kmart

Schlumberger

THOMSON

TOSHIBA

DELHAIZE

BOEING

lastminute.com

Corporate Express
A Buhmann Company

SIGMA-ALDRICH

L3
communications

FASTENAL
INDUSTRIAL & CONSTRUCTION SUPPLIES

Whirlpool

Weatherford

REALOGY

Endeca Delivers Business Results

Sample Customers



Endeca for
Intranet



Endeca for
Knowledge
Management



Endeca for
Human Capital
Management



Endeca for
Knowledge
Management



Endeca for
Intranet

Representative Metrics

Content downloads increased 400%

Intranet usage increased 30%,
call center volume decreased 20%

Intranet usage increased 175%

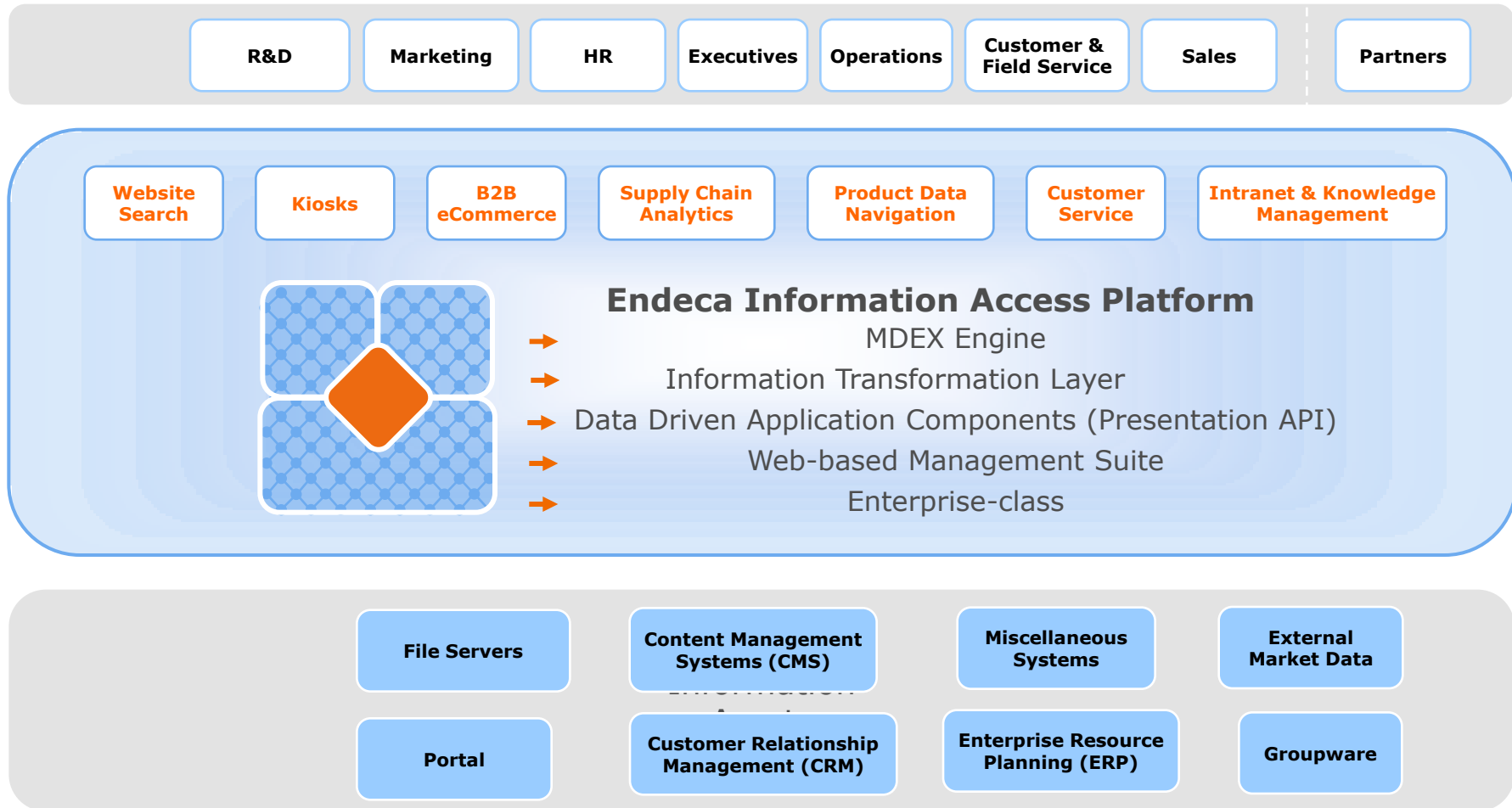
Saved \$500M through better
consultant staffing analysis

Significant employee efficiency
and productivity improvements

“The solution influences decisions
that generate \$100’s of millions in
revenue.”



Adaptivity closes the Information Access Gap



Endeca's unique architecture: **Adaptivity**

Adaptivity is
the **dynamic summarization** of
erratic data and content
in the current view

Case Study: Arabic in the Intel Community

Search: Within current results Dim Search Text Search In Field:

Filter By

person_org_loc

- Translated LOCATION [refine]
 - Ash Sharq Al Awsaq* (11)
 - Al `Irāq* (11)
 - Baghdād* (9)
 - Al Amrīkiyah* (8)
 - `Irāq* (8)
 - More...
- Translated ORGANIZATION [refine]
 - al-Hukumah al-'Iraqiyyah* (6)
 - al-Jaysh al-Amriki* (5)
 - al-Tayyar al-Sadri* (4)
 - Tanzim al-Qa'idah* (3)
 - Hizballah* (3)
 - More...
- Translated PERSON [refine]
 - al-Malikay* (7)
 - Nuri al-Malikay* (7)
 - Muqtada al-Sadr* (3)
 - Jurj Bush* (2)
 - Bush* (2)
 - More...

Breadcrumbs

[x]Text Search: All: المالكي

[x]Clear All Filters

Record Table **Record Detail**

Results 1-10 of 11 Results per Page: 10 | 20 | 30 Page: 1 of 2 >

		Document Title
		[x] ▾ ▲
1	View Record	العراق يتبرع بـ 35 مليون دولار إلى لبنان
2	View Record	مسؤول عسكري أمريكي: الزرقاوي باتن
3	View Record	الصحف العربية: عراق "الجنت" وحكومة "حملن"
4	View Record	القاعدة تؤكد مقتل الزرقاوي وتوعد بمواصلة الهجمات
5	View Record	بغداد: العثور على 547 جثة مجهولة منذ بداية مايو
6	View Record	صحف: غنيسن يهدد بالبقاء طويلا بالمنطقة وحاملة طائرات لمراقبة إيران
7	View Record	إيران تندد باستراتيجية بوش الجديدة بالعراق
8	View Record	مقتل أربعة جنود أمريكيين في الأندلس
9	View Record	الزعيم الشيعي مقتدى الصدر يظهر لأول مرة منذ شهرين بالكوفة
10	View Record	بريطانيا تعلن اختطاف خمسة من مواطنيها في العراق

Results 1-10 of 11 Results per Page: 10 | 20 | 30 Page: 1 of 2 >

Location Cloud **Person Cloud** **Organization Cloud** **Translated Cloud**

al-Malikay

Nuri al-Malikay

Muqtada al-Sadr Jurj Bush Bush

Saddam Husayn al-Zarqawi

Abaw Mus'ab al-Zarqawi

Mahmud 'Abbas 'Abbas

Fu'ad al-Sinyurah Isma'il Haniyyah

Byrts Usamah Bin-Ladin Allah Zarqawi 'Abdallah Musa 'Amru Musa Ulmirt Nasrallah al-Asad Shar al-Asad al-Hariri 'Anan More...

Arabic in the Intelligence Community

محمود عباس لا يستبعد كل الخيارات المتاحة أمامه من أجل وقف دائرة العنف في الأراضي الفلسطينية بما فيها الانتفاضة الثانية. ويتحدث عباس في مقابلة من رامها قتل مهاجمون سبعة أشخاص في بغداد وأربعة في ديالى، فيما استهدف مسلحون رئيسي بلديتي سامراء العراقية. أعلنت السلطات العراقية عن مقتل سائر
اعتقلت السلطات الإندونيسية أخطر مطلوبين الإرهاب في جنوب شرقي آسيا خلال حملة دهم في وسط جزيرة
تراجعت أسعار النفط في البورصة الآسيوية الثلاثاء بعد ارتفاعها أكثر من دولار في وقت متأخر الاثنين مدفوعاً باستبعاد الأوبك أي تحرك لرفع سقف الإنتاج الحالي خلال اجتماع المنظمة المقرر في سبتمبر/أيلول. وفي غضون ذلك، تترقب الأسواق العالمية بيانات الاحتياط الأمريكي، التي تصدرها إدارة معلومات الطاقة الأمريكية الأربعاء، وسط توقعات بارتفاع المخزون للأسبوع السادس على التوالي، جراء ارتفاع إنتاج المصافي والإستيراد.

- Arabic source data
- Analysts not fluent
- Limited access to translators and native speakers
- What to translate?

**Prioritize using key terms:
person, place, and organization**

Data Mining: Entity Extraction

على حافة حرب

Al Iraq Lubnan
لبنان يتبرع بـ 35 مليون دولار إلى العراق

1802 (GMT+04:00) - 22/08/06



Baghdad Al Iraq
رغم أزميتها الإنسانية .. (CNN) العراق ، بغداد
الحالية، تبرعت الحكومة العراقية بـ 35 مليون دولار
Lubnan
كمساعدهات لهيئة الإغاثة العليا بلبنان ، في إشارة على
al-Hukumah al-Iraqiyyah
دعم الحكومة اللبنانية ، وتعبير عن مخاوف
Al Iraq
العراق من اتساع حجم الأزمة الإنسانية التي يعاني
منها المواطنين اللبنانيين
Salih Al Iraq
وأعلن النرويج نائب رئيس الوزراء العراقي ، برهم صالح ، حيث قال إنه اتصل هاتفيا لرئيس
Fu'ad al-Sayurah
الوزراء اللبناني، فؤاد السنجورة ، الاثنان ليلتفعا بالتمردات المالية
Salih
ونكر مسؤول بارز في مكتب صالح : " لا يمكن أن ندع أشقائنا اللبنانيين يعانون بدون مساعدة
Al Iraq
من أشقائهم العرب. لقد طلبت رئيس الوزراء (العراقي) ودائمه، المجتمع الدولي بضرورة بذل
Lubnan
IRIN المزيد من الجهود لوقف العنف الذي انتشر في لبنان ، " وذلك حسب ما نقلته وكالة
Umsam al-Muttahidah
بالأمم المتحدة التابعة لخدمات أبناء ومعلومات عمليات الإغاثة الإنسانية



Persons

⇒ *Fuad al-Sinyurah*

⇒ *Salih*

Organizations

⇒ *Al-Hukumah al-Iraqiyyah*

Locations

⇒ *Al Iraq*

⇒ *Baghdad*

⇒ *Lubnan*

Data Mining: Name Similarity Indexing

- Linguistic analysis of name similarity
 - ⇒ *High likelihood*: "Al-Malikay" & "Nuri al-Malikay"
 - ⇒ *Possible*: "Mahmud Abbas" & "Abbas"
 - ⇒ *Unlikely*: "Fu'ad al-Sinyurah" & "Amru Musa"
- Enables fuzzy expansion of name searches
 - ⇒ Type "maliki" and search for
 - المالكي
 - al-Malikay
 - Nuri al-Malikay

Information Access: Fuzzy Keyword Search

Search:



Breadcrumbs

[X] Text Search: All: **الْقَذَافِي** OR **قَذَافِي**

[X] Clear All Filters

- Keyword search
- Stemming / inflection
- Thesaurus
- Spelling correction
- Wildcard
- Phrase interpretation
- DYM suggestions
- ...

Information Access: Navigation and Visualization

Filter By

person_org_loc

Translated LOCATION [refine]

Ash Sharq Al Awsaṭ (349)
Al Amrīkīyah (173)
Al Amrīkī (159)
Al Wilāyāt Al Muttahidah (158)
Wāshīnḡun (118)
More...

Translated ORGANIZATION [refine]

Tanzim al-Qa'idah (74)
al-Umam al-Muttahidah (61)
al-Qa'idah (59)
al-Barlaman (37)
Hizballah (33)
More...

Translated PERSON [refine]

Jurj Bush (78)
Bush (64)
Usamah Bin-Ladin (36)
al-Zarqawi (32)
Mu'ammār al-Qadhafi (17)
More...

Location Cloud

Person Cloud

Organization Cloud

Translated Cloud

Jurj Bush

Bush

Usamah Bin-Ladin

al-Zarqawi Saddam

Rayis Mu'ammār al-Qadhafi

al-Qadhafi Zarqawi

Kunduliza Rayis Ayman al-Zawahiri

Musa Abumus'ab al-Zarqawi

'Abdallah al-Thani 'Arafat al-Zarqawi Li Ibn-Ladin

al-Zarqawi Ra's Salih 'Ali Salih al-Nabhan

Nur-al-Din Muhammad Twb 'Abdallah Muhammad Salah

'Abd-al-'Aziz 'Udih Qadhafi Abubakr JANJAny

Muhammad 'Ali Hamadi Rubirt Stythm

Jamal Ahmad Muhammad 'Ali al-Badawi *More...*

Information Access: Transliteration in Context

Record Table

Record Detail

Results 1-10 of 11 Results per Page: 10 | 20 | 30 Page: 1 2 >

		Document Title
		[x] ▾ ▲
1	View Record	العراق يتبرع بـ 35 مليون دولار إلى لبنان
2	View Record	مسؤول عسكري أمريكي: الزرقاوي بائس
3	View Record	الصحف العربية: عراق "الجتت" وحكومة "حملة"
4	View Record	القاعدة تؤكد مقتل الزرقاوي وتوعد بمواصلة الهجمات
5	View Record	بغداد: العثور على 547 جثة مجهولة منذ بداية مايو
6	View Record	صحف: غيتس يهدد بالبقاء طويلا بالمنطقة وحملة طائرات لمراقبة إيران
7	View Record	إيران تدد باستراتيجية بوش الجديدة بالعراق
8	View Record	مقتل أربعة جنود أمريكيين في الأتارب
9	View Record	الزعيم الشيعي مقتدى الصدر يظهر لأول مرة منذ شهرين بالكوفة
10	View Record	بريطانيا تملن اختطاف خمسة من مواطنيها في العراق

Results 1-10 of 11 Results per Page: 10 | 20 | 30 Page: 1 2 >



على حافة حرب

Al 'Irāq Lubnān
العراق يتبرع بـ 35 مليون دولار إلى لبنان

1802 (GMT+04:00) - 22/08/06



Baghdād Al 'Irāq
رغم أزمته الإنسانية -- (CNN) العراق ، بغداد

al-Hukumah al-'Iraqiyyah
الحالية، تبرعت الحكومة العراقية بـ 35 مليون دولار

Lubnān
كمساعادت لهيئة الإغاثة العليا بلبنان ، في إشارة على

al-Hukumah al-Lubnaniyyah
دعم الحكومة اللبنانية ، وتعبير عن مخاوف

Al 'Irāq
العراق من اتساع حجم الأزمة الإنسانية التي يعاني

منها المزارعين اللبنانيين

Salih Al 'Irāqī
وأعلن التبرع نائب رئيس الوزراء العراقي ، برهم صالح ، حيث قال إنه اتصل هاتفيا لرئيس

Fu'ad al-Sinuyrah
الوزراء اللبناني، فؤاد السنيورة ، الاثنين ليبلغه بالترقيات المالية

Salih
وذكر مسؤول بارز في مكتب صالح : " لا يمكن أن ندع أشقائنا اللبنانيين يعانون بدون مساعدة

Al 'Irāqī
من أشقائهم العرب. لقد طالب رئيس الوزراء (العراقي) ونائبه، المجتمع الدولي بضرورة بذل

Lubnān
IRIN المزيد من الجهود لوقف العنف الذي انتشر في لبنان ،" وذلك حسب ما نقلته وكالة

Umam al-Muttahidah
بالأمم المتحدة التابعة لخدمات أنباء ومعلومات عمليات الإغاثة الإنسانية