I18n/L10 of a Search Engine

Internationalization and Unicode Conference

March 2006 | Tuoc V. Luong

Agenda

Ask.com Introduction

- Ask.com Search
- Components of a Search Engine

Technical Issues

- Software and Indices Data

Product Issues

- Search within Language, Country or World

Development Process Issues

Ask.com Introduction

Ask.com Profile

- # 7 US & Global web property
- # 5 ranked Search property
- 25% reach – US audience
- 40 million domestics & 132 million global unique users
- 6.4% share of US searches
- A division of IAC/Interactive Corp.

History of Search at Ask

Questions and Answers

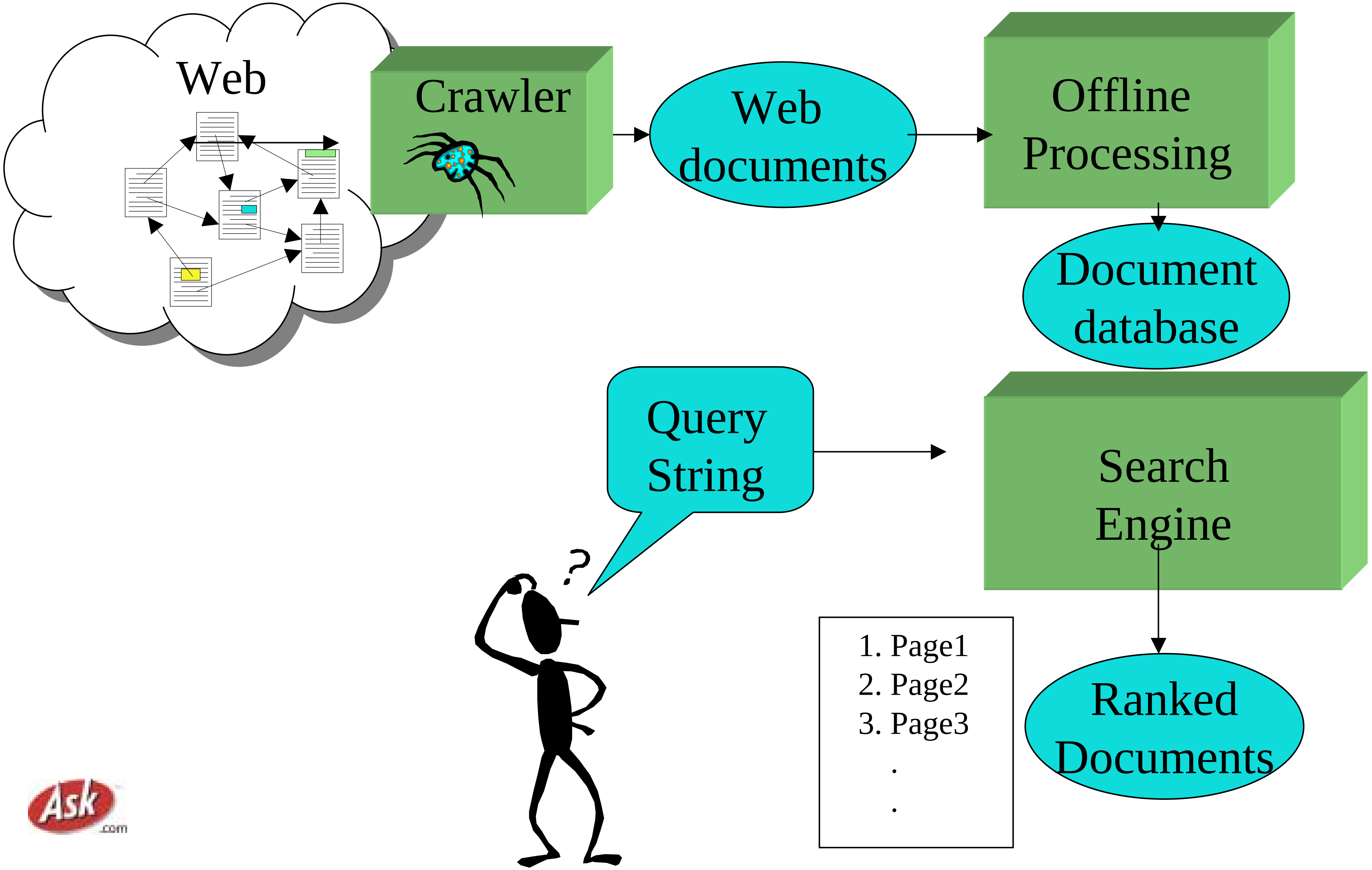
- Technology focus was on Questions (NLP)
- Historically answers were provided editorially

Bought Teoma (Sept. 2001)

- Foundation of current Ask.com search
- Hub and authority (subject specific)
- Originally very U.S. Centric
- Today – internationalized and supporting

  English (USA & UK), Japanese, FIGS, and Dutch

Infrastructure for Web Search



Web

Crawler

Web documents

Offline Processing

Document database

Query String

Search Engine

1. Page1
2. Page2
3. Page3
.
.

Ranked Documents

Technical Issues

Technical Issues

- Document Language Identification
- Segmentation and Normalization
- User Language Identification
- Content Classification by Country, Language, …etc
- Geographical Distances

Language and Character Set Issues

- **Documents with Low Text**
  - Difficult to identify language
- **Number of Languages is Very Large**
  - But unevenly distributed on the Web
- **Documents with Mixed Languages**
- **Documents with Mixed Character Sets**
  - increasingly more common with Blogs (replies)
- **Documents are Converted to UTF8**
  - Detect language and character set after download

Many Language but One Engine

- Learning as We Go Along
- Segmentation – Japanese and Chinese
- Accents – French and Spanish
- Compounding – German and Dutch
- Spelling Variations – All
- How to Architect and Organize for Languages?

Standard Model of Language

- **Normalize Text**
  - Reduce orthographic variations (facilitate recall)
- **Tokenize Text**
  - Divide text into stream of tokens (searchable)
- **Spelling Variants, Compounding and other Features**
  - Device language-independent representation

Can We Be Too Language Specific?

- Words are Borrowed from one Language to Another
- Mixed Language Documents
- How do we Apply Language Specific Rules here?
- Generalize Language Features across Languages
- Easier said than done in practice.

Multi-Language Search Engine?

- **Most Users Understand One Language Well**
  - Information in other languages not as useful
- **Engine Should Know What Languages User Wants**
- **Search for Languages that the User Understands**
- **Translate to Access Documents in other Languages**

Determine Language That User Wants

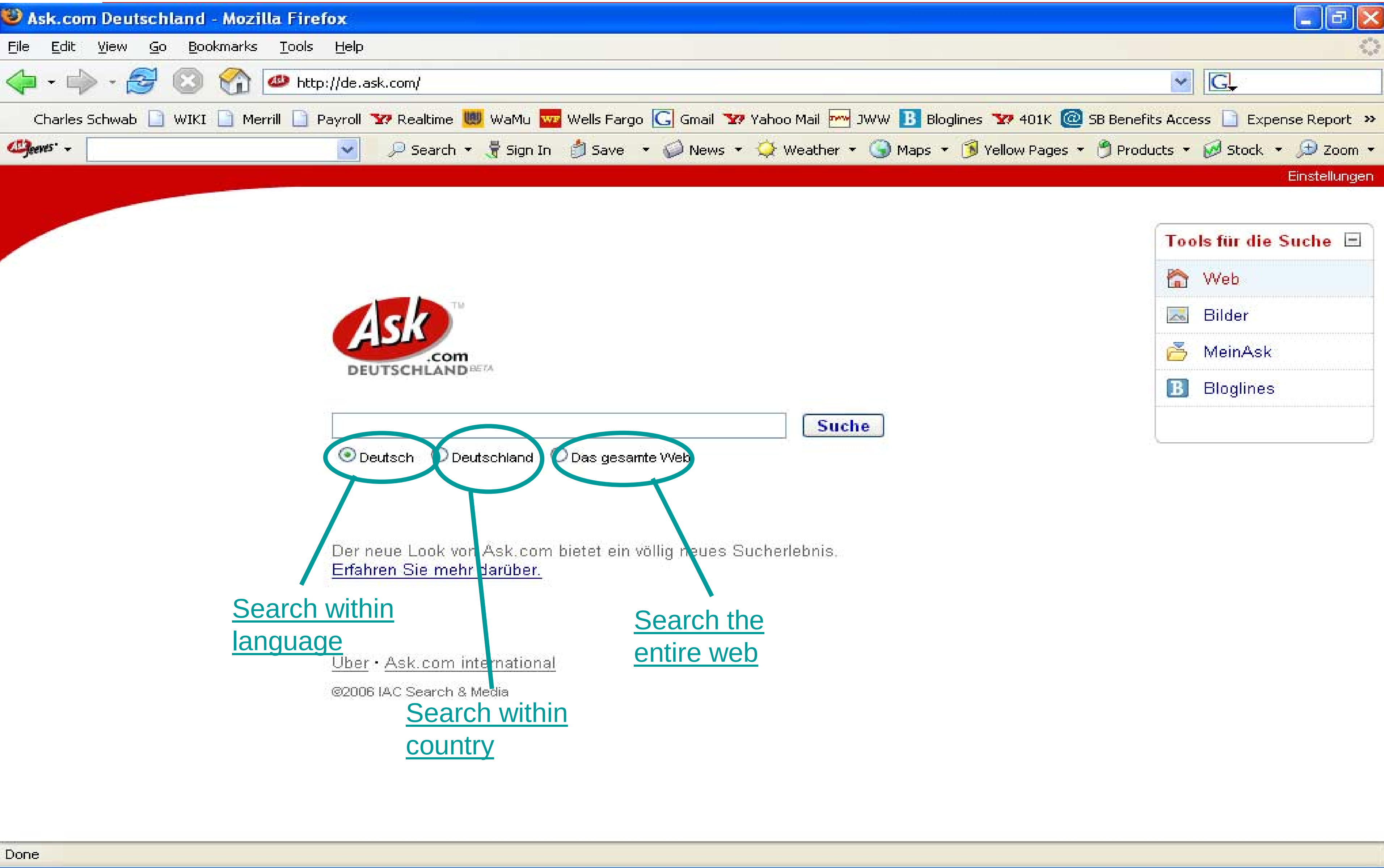
- Operating system and browser settings?
- Country where user is located (IP address)?
- Web site they have visited (ask.com, ask.jp)?
- Explicit preferences (advanced page)?
- Analyze the query itself?

Geographical Location

- **Internet Latency across Continents is Substantial**
- **Search Engine Close to Users makes a Difference**
  - Can be very expensive to implement
- **Dividing up Documents Works Sometimes**
  - Can be difficult/expensive for world search

Product Issues

Search w/ Language, Country and World



Ask.com Deutschland - Mozilla Firefox

http://de.ask.com/

○ Deutsch ○ Deutschland ○ Das gesamte Web

Der neue Look von Ask.com bietet ein völlig neues Sucherlebnis.
Erfahren Sie mehr darüber.

Über · Ask.com international

©2006 IAC Search & Media

Search within language

Search within country

Search the entire web

Classifying The Web

		Country					
		US	France	Canada	UK	Belgium	Australia
LANG	English			🟥			
	French	🟦	🟦	🟪	🟦	🟦	🟦
	Chinese			🟥			
	German			🟥			
	Japanese			🟥			
	…			🟥			

Consistency Across Products

- **Many Teams Working on Many Different Products**
- **Web Search, News Search, Image Search, … etc**
- **Consistency in Handling Language ID, Text Normalization, … etc**
- **Libraries with Key International Functionality**
    - Shared / leveraged throughout technology teams

Software Development Issues

Process issues

- Transition
- Setting goals for new projects
- Team selection
- Team organization
- Business entity organization

Retroactively I18n Search Engine

- **Massive Effort to I18n Entire Search Engine Code Base**
  - Meanwhile continue delivering new features and scale
- **Japan was our First Targeted Delivery Outside of USA/UK**
- **Convert the Existing Search Engine to UTF-8 first?**
- **Diverge and Create a Separate I18n/Japanese Version?**
- **Diverge First, then Merge Back Together?**
  - Ultimately that’s what we did
  - Merging back to the main engine was harder than launching the Japanese engine

Team Selection

- **Software Engineers to Create the Product**
  - Core engineering + I18n specific specialists
- **Content Editorial Staff in each Language**
  - Evaluate relevance, remove undesirable contents, identify issues
- **Engineering Team Can’t Tell Quality of Product**
  - Are the local language/country results relevant?
- **The Teams Don’t Communicate That Readily**
  - Language facility; technical/non-technical people

Team Location

- **Content Staff are Native Speakers**
  - Need to stay immersed in their culture
  - Therefore located in or near the target country

- **I18n Engineers Work Side by Side with Core Eng**

- **Lots of Phone Meetings at Strange Hours**
  - Not the most productive way to work but necessary for a global information company

Summary

Internationalization is Best Done from Beginning

- Not always possible and thus expensive to retro
- It was quite an expensive endeavor but necessary

Hope You Got a Good View into the Issues

- Technical Issues
- Product Issues
- Process Issues

Do it from the START!!!