Query Expansion from Wikipedia and Topic Web Crawler on CLIR

Meng-Chun Lin, Ming-Xiang Li, Chih-Chuan Hsu, Shih-Hung Wu

Proceedings of NTCIR-8 Workshop Meeting, June, 2010

Adviser : Prof. Shih-Hung Wu

Reporter : Meng-Chun Lin

Outline

- Introduction
- Translation Methods
  - Wikipedia Translation
  - Online Translation Website
- Query Expansion Methods
  - Thesaurus - Wikipedia
  - Pseudo Relevance Feedback
- Topic web crawler
- Experiment Result
- Conclusions

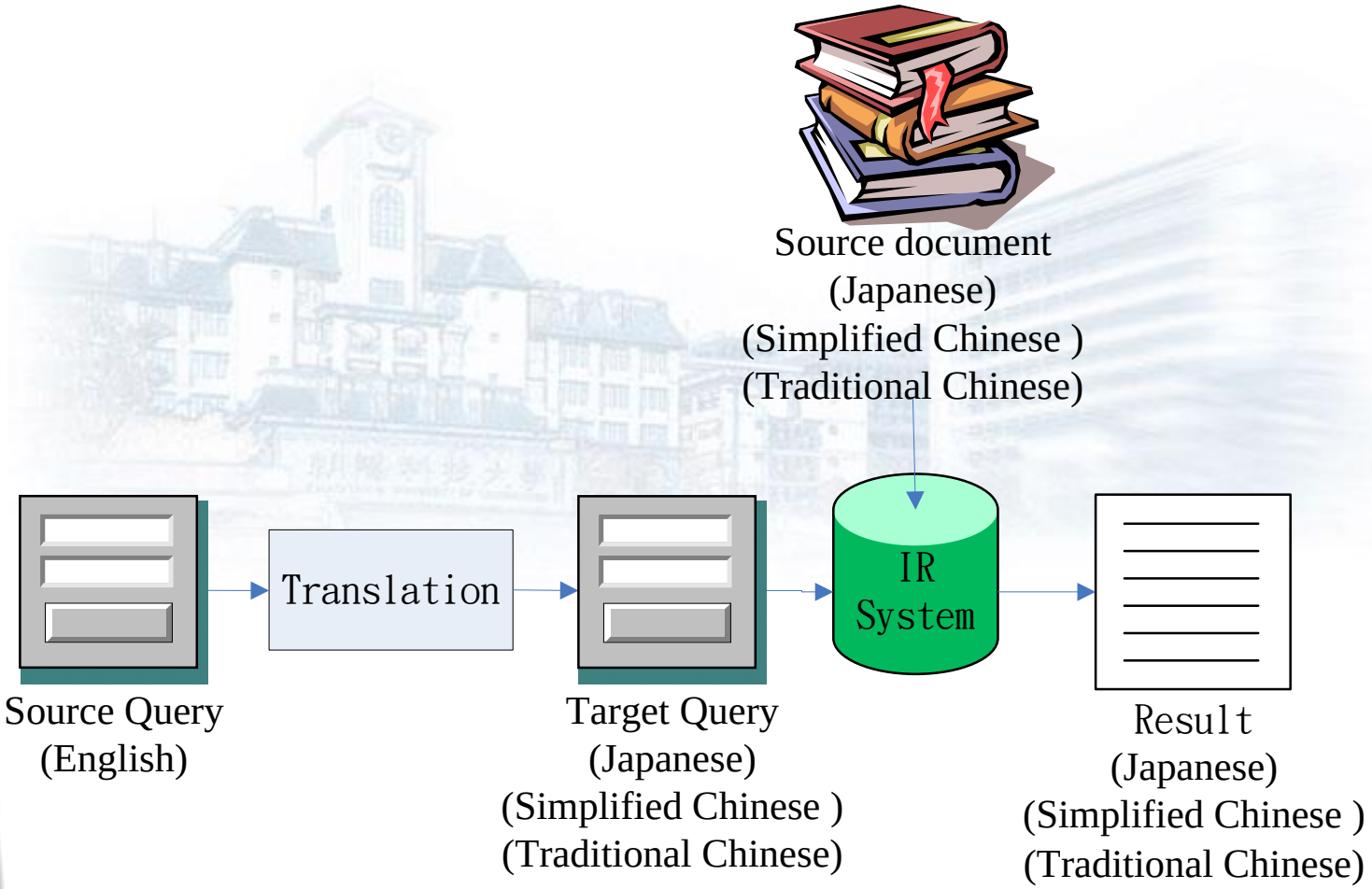
Introduction

- In this paper, we report various strategies for query expansion (QE) in the NTCIR-8 IR4QA subtask. We submit the results of twelve runs from the formal run.

- The results of twelve runs include cross-language information retrieval from English to traditional Chinese, from English to simplified Chinese, and from English to Japanese in the official T-run, D-run and DN-run.

Introduction

2. Query translation :



Introduction

- In our previous works, Su et al. [2007] adopted online translation website services as a fixed dictionary and Wikipedia as a live dictionary to translate query terms. Their method can translate Out Of vocabulary (OOV) terms efficiently.

- Lin et al. [2008] purposed a method that combines OKAPI BM25 and Wikipedia anchor texts for query expansion.

Introduction

- In this paper, we combine Su's and Lin's methods in our system. Then we add more QE from Wikipedia and the result of QA analysis.

- In the additional runs, we use a topic web crawler to get more related web pages and extract more keywords to be the candidates of QE.

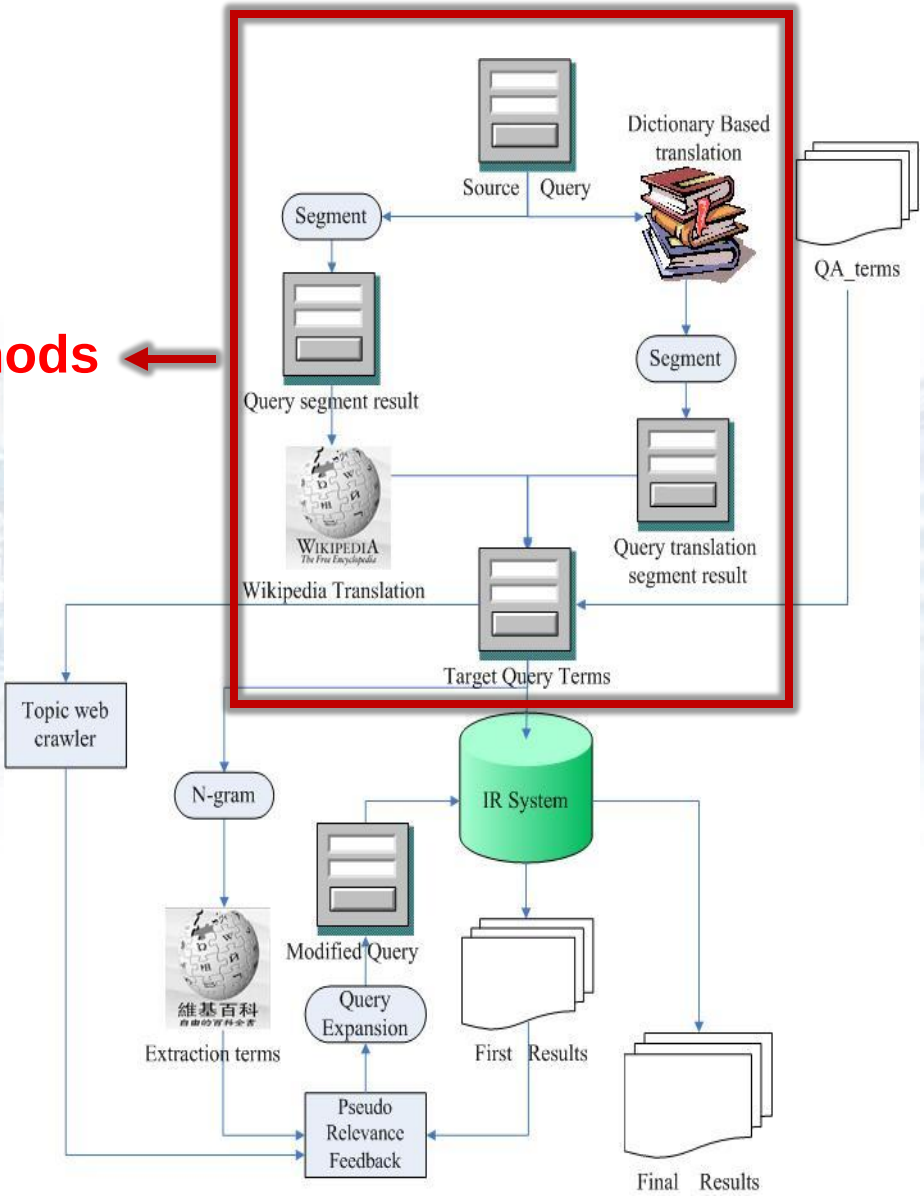
Introduction

- Finally, we make use of Wikipedia, a good information resource, and topic crawler, to extract more keywords to be the candidates of QE, to improve our precision in CLIR.

Translation Methods

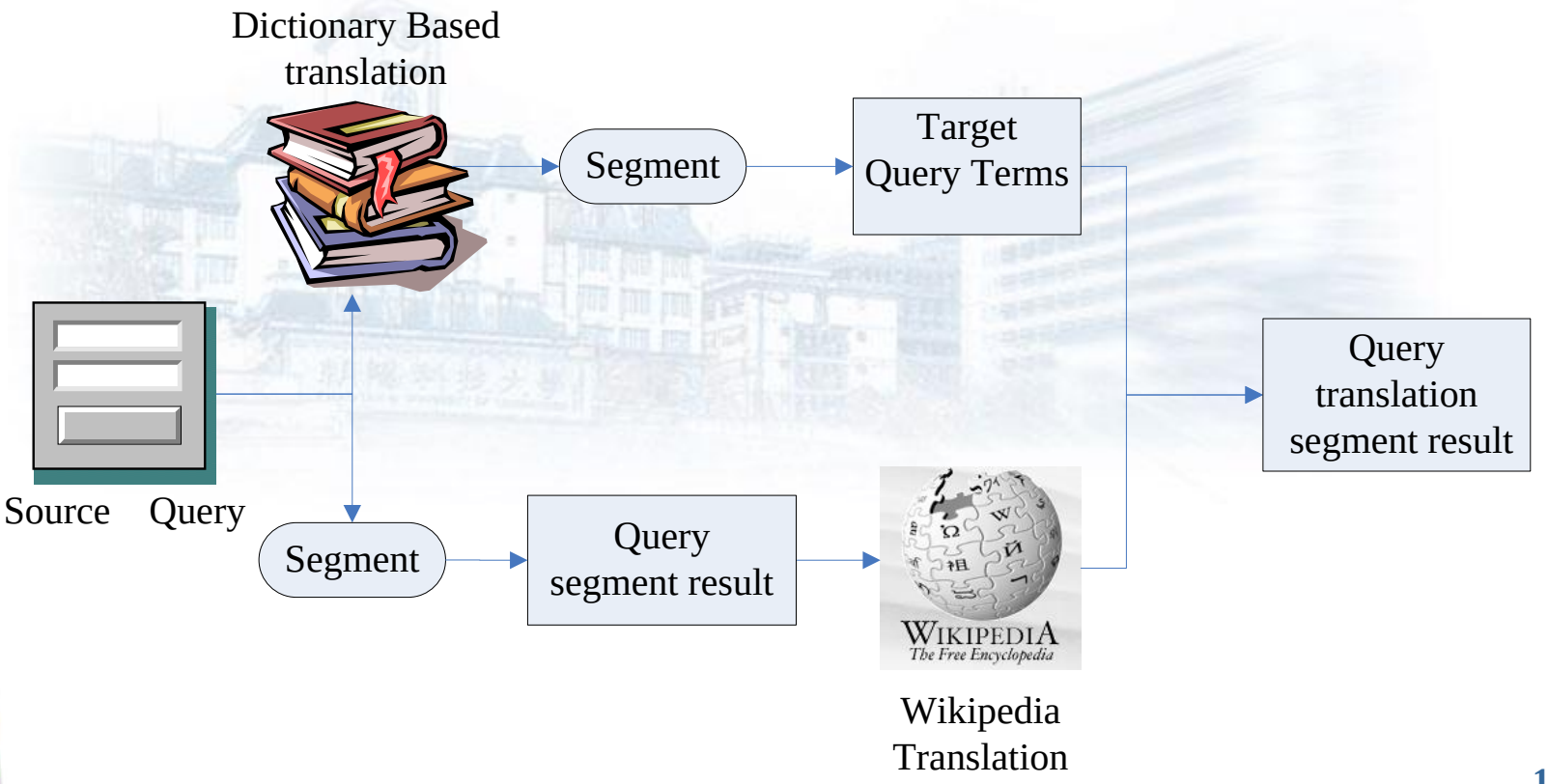
Architecture of retrieval system

Translation Methods



Our Translation System

- Translation system flow chart:



Wikipedia Translation

- Wikipedia is a multilingual encyclopedia on the web and is composed and edited by volunteers all over the world. Anyone can edit or create new articles.
- Total has more then 15 million articles in 270 languages. The numbers of articles still grow up.
  - Retrieved from May, 2010

Languages	Articles
English	3,285,662
Japanese	674,217
Chinese	307,698
29 languages	>100,000

Wikipedia Translation

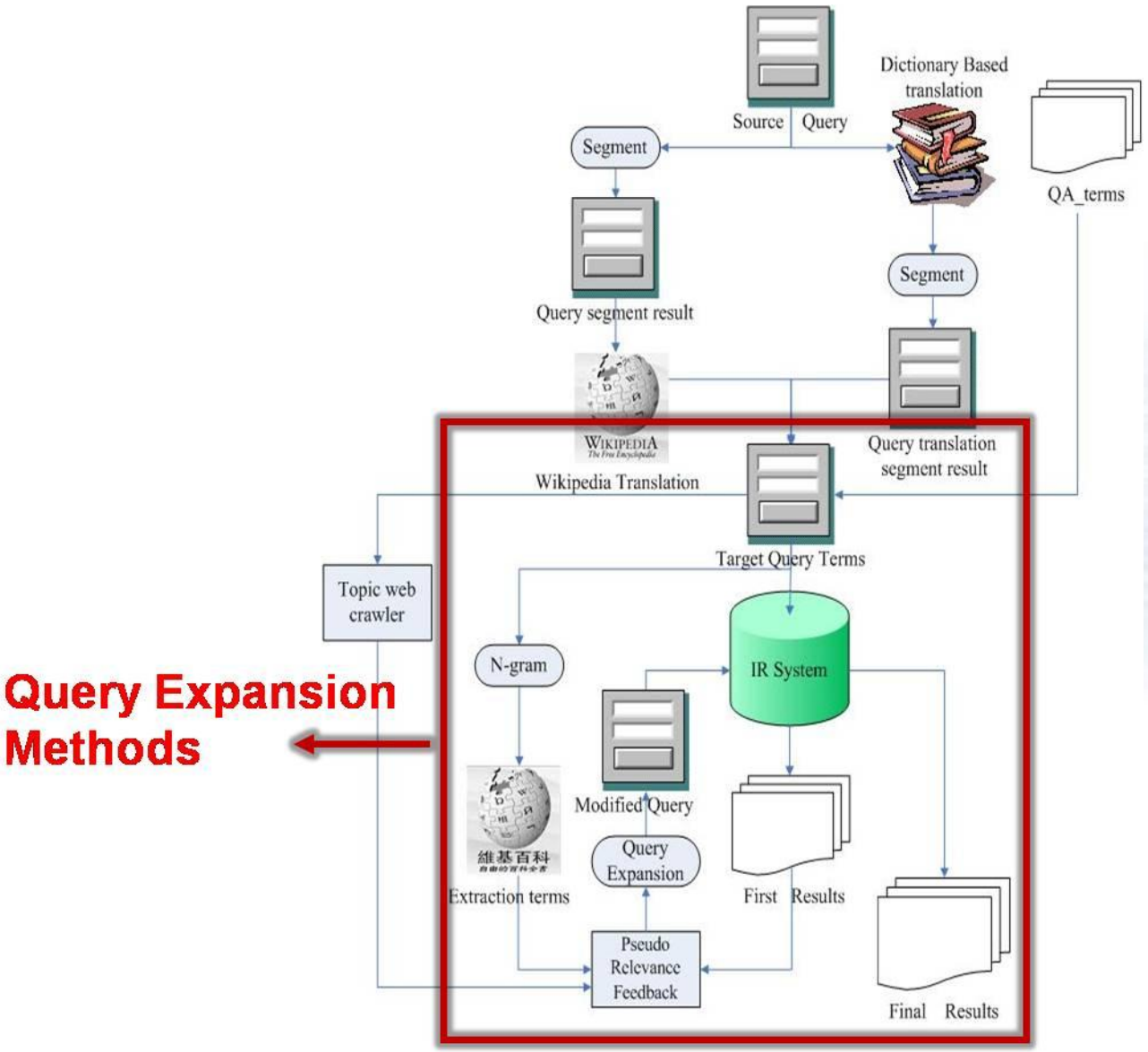
- Wikipedia translation method:

Online Translation Website

- Google translation:

Query Expansion Methods

Architecture of retrieval system

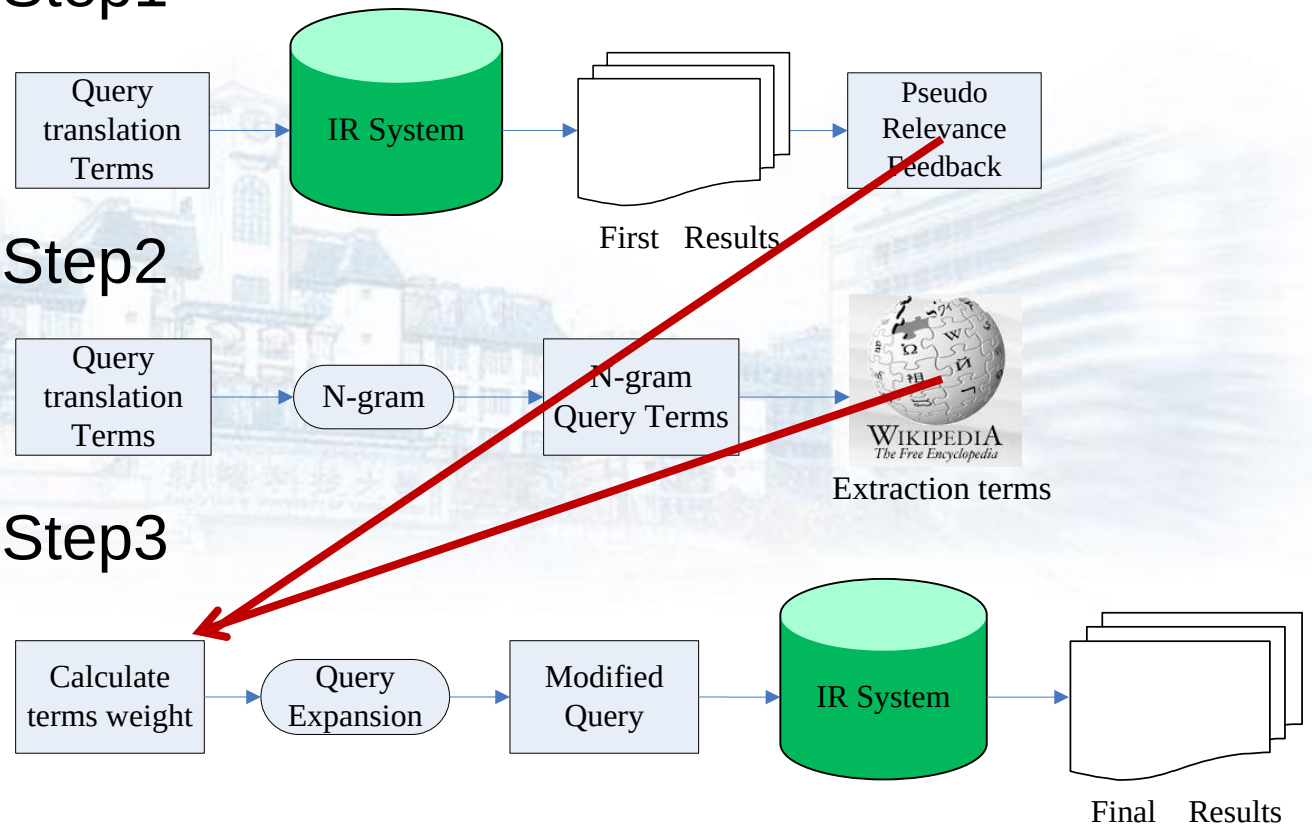


Query Expansion

- Query expansion is an important technology in IR systems since it can increase recall value.
- There are two major approaches:
  - Thesaurus
  - Pseudo relevance feedback
- We combine these two methods in our experiments by treating Wikipedia as a kind of thesaurus.

Our Retrieval System

- Retrieval system flow chart:
  - Step1
  - Step2
  - Step3



Wikipedia Query Expansion

- In Wikipedia, every entry has links to related entries or other relevant web pages on other websites.
- The anchor texts of the hyperlinks must be related terms. Therefore, we treat these anchor texts as candidates for query expansion.

Entry

Log in / create account

article | discussion

Support Wikipedia: a non-profit project. Donate Now »

[Expand]

Nepal

From Wikipedia, the free encyclopedia

Coordinates: 26°32 N 86°44 E

This article **may require copy-editing for grammar, style, cohesion, tone or spelling**. You can assist by editing it now. A how-to guide is available. *(February 2008)*

Nepal (Nepali: नेपाल [ne'pa:l] (help·info)) is a landlocked country in South Asia. It is bordered by the People's Republic of China to the north and by India to the south, east and west. The Himalaya mountain range runs across Nepal's northern and western parts, and eight o the world's ten highest mountains, including the highest, Mount Everest, lie within its borders.

Historically, what is now Nepal was made up of many small kingdoms. The modern state was formed with the Unification of Nepal by Prithvi Narayan Shah on December 21, 1768. Until 2006, Nepal was a kingdom. Nepal is now a federal democratic republic.[3] Its recent history has involved struggles for democratic government with periods of direct monarchic rule. From 1996 until 2006, there was a Civil War between government forces and guerrillas of the Communist Party of Nepal (Maoist).

On December 28, 2007, the Interim Parliament passed a bill and declared Nepal to be a *Federal Democratic Republic*. The first meeting o the Constituent Assembly officially implemented that declaration on May 28, 2008.

Nepal is a multi-cultural, multi-linguistic and multi-religious country. For a relatively small country, Nepal has a diverse landscape, ranging from the humid Terai plains in the south to the mountainous Himalayas in the north, ism is practiced by a large[specify] majority of the people, but the country also has a stror Buddha Siddhartha Gautama is located in the Terai, one of the three regions of Nepa

Short Content

The capital Kathmandu is the largest city in the country. The official language is Nepali and the state currency is the Nepalese Rupee (NPR).

Nepal's flag is the only national flag in the world that is non-quadrilateral in shape. The blue border on the flag of Nepal signifies peace. The red in the flag stands for victory in war or courage, and is also color of the rhododendron, the national flower of Nepal. While the curved moon on the flag is a symbol of the peaceful and calm nature of Nepalese, the sun represents the aggressiveness of Nepalese warriors.

Long Content

Contents [show]

Etymology

Link & Anchor Text

Nepal Bhasa

The word "Nepal" is believed by scholars to be derived from the word "Nepa:" which refers to the Newar Kingdom, the present day Kathmandu Valley. With Sanskritization, the Newar word Nepa became Nepal.[4] The Newars of present day Nepal, refer to all the inhabitants of Kathmandu valley and its peripheries (called "Nepa:") before the advent of Shah dynasty. The Nepal Sambat calendar, named after this Newar kingdom and devised 1100 years ago, is a national calendar used in Nepal and testifies to its antiquity.

Ne Muni

संघीय लोकतान्त्रिक गणतन्त्र नेपाल
Sanghiya Loktāntrik Ganatantra Nepāl
Federal Democratic Republic of Nepal

Flag

Emblem

Motto: "Mother and Motherland are Greater than Heaven"

Anthem: "Sayaun Thunga Phool Ka"

Capital (and largest city)	Kathmandu (Nepali: काठमांडौं) 27°42'N 85°19'E
Official languages	Nepali[1]
Recognised regional languages	Maithili, Nepal Bhasa, Bhojpuri, Tharu, Gurung, Tamang, Magar, Awadhi, Sherpa, Kiranti and another 100 different indigenous languages.
Demonym	Nepali

navigation
- Main page
- Contents
- Featured content
- Current events
- Random article

search
Go | Search

interaction
- About Wikipedia
- Community portal
- Recent changes
- Contact Wikipedia
- Donate to Wikipedia
- Help

toolbox
- What links here
- Related changes
- Upload file
- Special pages
- Printable version
- Permanent link
- Cite this page

languages
- Afrikaans
- Alemannisch
- العربية
- Aragonés
- Arpetan
- Asturianu
- Azərbaycan
- বাংলা
- Bân-lâm-gú
- Беларуская

ウィキペディア
フリー百科事典

ログインまたはアカウント作成

本文 ノート

非営利プロジェクトをご支援ください。 今すぐに寄付 »

[表示]

出典: フリー百科事典『ウィキペディア(Wikipedia)』

この項目では南アジアの國ネパールについて記述しています。ネパールの政治家についてはマダブ・クマル・ネパールをご覧ください。

この項目は現在進行中の事象を扱っておりますが、ウィキペディアはニュース速報ではありません。性急な編集をせず検証可能な事実を確認し、正確な記述を心がけてください。またウィキニュースへの投稿も検討してみてください。なお、この内容は不特定多数のボランティアにより自由に編集されていることを踏まえ、自身の安全利害に関わる情報は自己責任でご判断ください。

ネパールは南アジアの連邦民主共和國(2008年に王制廃止)。

南にインド、北に中國チベット自治區を接する東西に細長い内陸國である。國土は世界最高地點エベレスト(サガルマータ)を含むヒマラヤ山脈および中央部丘陵地帯と、南部のタライ平原から成る。ヒマラヤ登山の玄関口としての役割を果たしている。

多民族・多言語國家(インド・アーリア系の民族と、チベット・ミャンマー系民族)であり、民族とカーストが複雑に関係し合っている。また、宗教もヒンドゥー教(元國教)、仏教、アニミズム等とその習合が混在する。

2008年5月、制憲議會が発足したが、新政権は3か月にしてようやく8月31日、難航の末プラチャンダ内閣が本格的に発足した。1996年から2006年までネパール共産党毛沢東主義派と政府の間で内戦[illegible](UNMIN)を受けている。

経済[illegible]ヒマラヤ観光など[illegible]

目次 [表示]

國名

- 日本語表記:**ネパール連邦民主共和國**[1]。
- 公式の英語表記:**Federal Democratic Republic of Nepal**、略稱:*Republic of Nepal*[2]

正式名稱はネパール語のデバナガリ(デーヴァナーガリー)文字で**संघीय लोकतान्त्रिक गणतन्त्र नेपाल**、ラテン文字転寫表記は*saṅghīya loktāntrik gaṇatantra nepāl*。略稱**गणतन्त्र नेपाल**、ラテン文字転寫表記は*gaṇatantra nepāl*。通稱**नेपाल**、ラテン文字転寫表記は*Nepāl*。

國民 [編集]

ネパール政府は1958年に中央統計局(Central Bureau of Statistics)を設け、10年に一度國勢調査を行うほか、國民所得統計、農業センサスなども行っている。また、サンプル調査により、毎年人口推計を出している。

ネパール連邦民主共和國
संघीय लोकतान्त्रिक गणतन्त्र नेपाल

(國旗) (國章)

國の標語:*जननी जन्मभूमिश्च स्वर्गादपि गरियसि*
ラテン文字転寫:Janani Janmabhumishcha Swargadapi Gariyasi
(サンスクリット語:祖國は天國より素晴らしい)

國歌:國歌(題名無し)

公用語	ネパール語
首都	カトマンズ
最大の都市	カトマンズ
政府	

ナビゲーション
- メインページ
- コミュニティ・ポータル
- 最近の出來事
- 最近更新したページ
- おまかせ表示
- アップロード(ウィキメディア・コモンズ)
- ウィキペディアに関するお問い合わせ

ヘルプ
- ヘルプ
- 井戸端
- お知らせ
- バグの報告
- 寄付

検索
表示 検索

ツールボックス
- リンク元
- リンク先の更新状況
- 特別ページ
- 印刷用バージョン
- この版への固定リンク
- この項目を引用

他の言語
- Afrikaans
- Alemannisch
- Aragonés
- العربية
- Asturianu
- Azərbaycan
- Žemaitėška

维基百科
自由的百科全书

搜索
进入 搜索

导航
- 首页
- 分类索引
- 特色内容
- 新闻动态
- 最近更改
- 随机页面

帮助
- 帮助
- 社区
- 方针与指引
- 互助客栈
- 询问处
- 字词转换
- 联系我们
- 关于维基百科
- 资助维基百科

工具箱
- 链入页面
- 链出更改
- 上传文件
- 特殊页面
- 可打印版
- 永久链接
- 引用此文

其他语言
- Afrikaans
- Alemannisch
- Aragonés

登录 / 创建账户

条目 讨论 不转换 简体 繁体 大陆简体 港澳繁体 马新简体 台湾正体

Entry

请资助维基百科：

[显示]

尼泊尔

维基百科，自由的百科全书

Short Content

尼泊尔（尼泊尔语：नेपाल），正式名称为**尼泊尔联邦民主共和国**（संघीय लोकतान्त्रिक गणतन्त्रात्मक नेपाल），[1]为南亚山区内陆国家，位于喜玛拉雅山脉，北与中国西藏相接，其余三面与印度为邻。中尼边界的西藏圣山珠穆朗玛峰（尼泊尔称**萨加玛塔峰**）是全世界最高的山峰。由于取道尼泊尔上山比较轻松，不少游客都取道尼泊尔登山。

原国号为**尼泊尔王国**。2006年8月，议会通过一项宪法修正案，改国号为**尼泊尔**。2008年5月28日，制宪会议通过决议废除长达240年的王室成立共和制，成为一个联邦民主共和国，

目录 [显示]

历史

[编辑]

Long Content

主条目：尼泊尔历史

尼泊尔是亚洲的古国之一。古代尼泊尔境内有很多国家，在前6世纪，尼泊尔人就已在加德满都河谷一带定居。12世纪前，被印度的加纳克国王的兄[illegible]帕尔、阿希尔、吉拉迪、李查维等王朝。

Link & Anchor Text

藏缅语族的尼瓦尔人（Newari）被认为是加德满都谷地的原住民，但他们却并不全都是东方的蒙古人种，还有许多是雅利安人。17世纪是尼瓦尔人的“黄金时代”，马拉王朝统治下的尼泊尔是西藏和北印度平原之间极为重要的贸易枢纽。当时，加德满都、帕坦和芭克塔普各自为政，三个城市之间的竞争非常激烈。

17世纪中叶廓尔喀人兴起，在西部甘达基河沿岸建立了一个小王国(沙阿王朝的前身)，1768年，巴里斯威·那拉扬·沙阿（Prithvi Narayan Shah）统一了尼泊尔地区，结束了加德满都谷地三城分地割据的状态。尼泊尔语（Nepali）——西部地区的一种印欧语系语言，代替了尼瓦尔语成为官方语言。

时值清朝国势强盛，康熙时期西藏已经被清朝控制，廓尔喀则成为清朝的藩属，向清朝进贡。而英国在占领印度后，渐渐向北进发，经常侵略哲孟雄、不丹等小国。所以，廓尔喀和清朝一直保持着良好的宗藩关系，以遏制英国的侵略。但自清朝中叶国势衰弱，清廷被内忧外患困扰，无暇理会外藩。中华民国建立后，袁世凯曾想邀请尼泊尔（即廓尔喀）加入“五族共和”，但当时的尼泊尔已经受英国控制。[来源请求]

1791年英国与尼泊尔签订了一项掠夺性的“通商条约”。1815年英国与尼泊尔签订了“塞格里条约”，强迫尼泊尔把南部大片土地割给东印度公司，并要求尼泊尔在内政和对外贸易方面接受英国的监督。1846年，亲英的廓尔喀军人拉腊发动政变，夺得尼泊尔军政要职，国王的大权旁落，拉腊家族世袭首相。1923年英国承认尼泊尔的独立，并与尼泊尔签订了“永久和平条约”。

संघीय लोकतान्त्रिक गणतन्त्रात्मक नेपाल
尼泊尔联邦民主共和国

通称：尼泊尔

国旗

国徽

国家格言：梵文：जननी जन्मभूमिश्च स्वर्गादपि गरीयसी
母亲和祖国比天堂更宝贵

国歌：Sayaun Thunga Phool Ka

自然地理
（实际管辖区）

首都	加德满都
最大城市	加德满都

面积
- 国土面积：140,800平方公里（世界第94名）
- 水域率：2.8%

Entry

Short Content

Long Content

Link & Anchor Text

尼泊爾

維基百科，自由的百科全書

維基百科 自由的百科全書

搜索

進入 搜索

導航
- 首頁
- 分類索引
- 特色內容
- 現時事件
- 最近更改
- 隨機頁面

幫助
- 幫助
- 社區
- 方針與指引
- 互助客棧
- 詢問處
- 字詞轉換
- 聯繫我們
- 關於
- 資助維基百科

工具箱
- 鏈入頁面
- 鏈出更改
- 上傳檔案
- 特殊頁面
- 可列印版
- 永久連結
- 引用此文

其它語言
- Afrikaans
- Alemannisch
- Aragonés

尼泊爾（尼泊爾語：नेपाल），正式名稱爲**尼泊爾聯邦民主共和國**（संघीय लोकतान्त्रिक गणतन्त्रात्मक नेपाल），[1]爲南亞[...]區內陸國家，位於喜瑪拉雅山脈，北與中國西藏相接，其餘三面與印度爲鄰。中尼邊界的西藏聖山珠穆朗瑪峰（尼泊爾稱**薩加瑪塔峰**）是全世界最高的山峰。由於取道尼泊爾上山比較輕鬆，不少遊客都取道尼泊爾登山。

原國號爲**尼泊爾王國**。2006年8月，議會通過一項憲法修正案，改國號爲**尼泊爾**。2008年5月28日，制憲會議通過決議廢除長達240年的王室成立共和制，成爲一個聯邦民主共和國，目[...]

目錄 [顯示]

歷史

[編輯]

主條目：尼泊爾歷史

尼泊爾是亞洲的古國之一。古代尼泊爾境內有很多國家，在前6世紀，尼泊爾人就已在加德滿都河谷一帶定居。12世紀前，被印度的加納克國王的兄弟[...]爾、阿希爾、吉拉迪、李查維等王朝。

藏緬語族的尼瓦爾人（Newari）被認爲是加德滿都谷地的原住民，但他們卻並不全都是東方的蒙古人種，還有許多是雅利安人。17世紀是尼瓦爾人的「黃金時代」，馬拉王朝統治下的尼泊爾是西藏和北印度平原之間極爲重要的貿易樞紐。當時，加德滿都、帕坦和芭克塔普各自爲政，三個城市之間的競爭非常激烈。

17世紀中葉廓爾喀人興起，在西部甘達基河沿岸建立了一個小王國(沙阿王朝的前身)，1768年，巴裡斯威·那拉揚·沙阿（Prithvi Narayan Shah）統一了尼泊爾地區，結束了加德滿都谷地三城分地割據的狀態。尼泊爾語（Nepali）——西部地區的一種印歐語系語言，代替了尼瓦爾語成爲官方語言。

時值清朝國勢強盛，康熙時期西藏已經被清朝控制，廓爾喀則成爲清朝的藩屬，向清朝進貢。而英國在佔領印度後，漸漸向北進發，經常侵略哲孟雄、不丹等小國。所以，廓爾喀和清朝一直保持著良好的宗藩關係，以遏制英國的侵略。但自清朝中葉國勢衰弱，清廷被內憂外患困擾，無暇理會外藩。中華民國建立後，袁世凱曾想邀請尼泊爾（即廓爾喀）加入「五族共和」，但當時的尼泊爾已經受英國控制。[來源請求]

1791年英國與尼泊爾簽訂了一項掠奪性的「通商條約」。1815年英國與尼泊爾簽訂了「塞格里條約」，強迫尼泊爾把南部大片土地割給東印度公司，並要求尼泊爾在內政和對外貿易方面接受英國的監督。1846年，親英的廓爾喀軍人拉臘發動政變，奪得尼泊爾軍政要職，國王的大權旁落，拉臘家族世襲首相。1923年英國承認尼泊爾的獨立，並與尼泊爾簽訂了「永久和平條約」。

संघीय लोकतान्त्रिक गणतन्त्रात्मक नेपाल
尼泊爾聯邦民主共和國

通稱：尼泊爾

國旗

國徽

國家格言：梵文：जननी जन्मभूमिश्च स्वर्गादपि गरीयसी
母親和祖國比天堂更寶貴

國歌：Sayaun Thunga Phool Ka

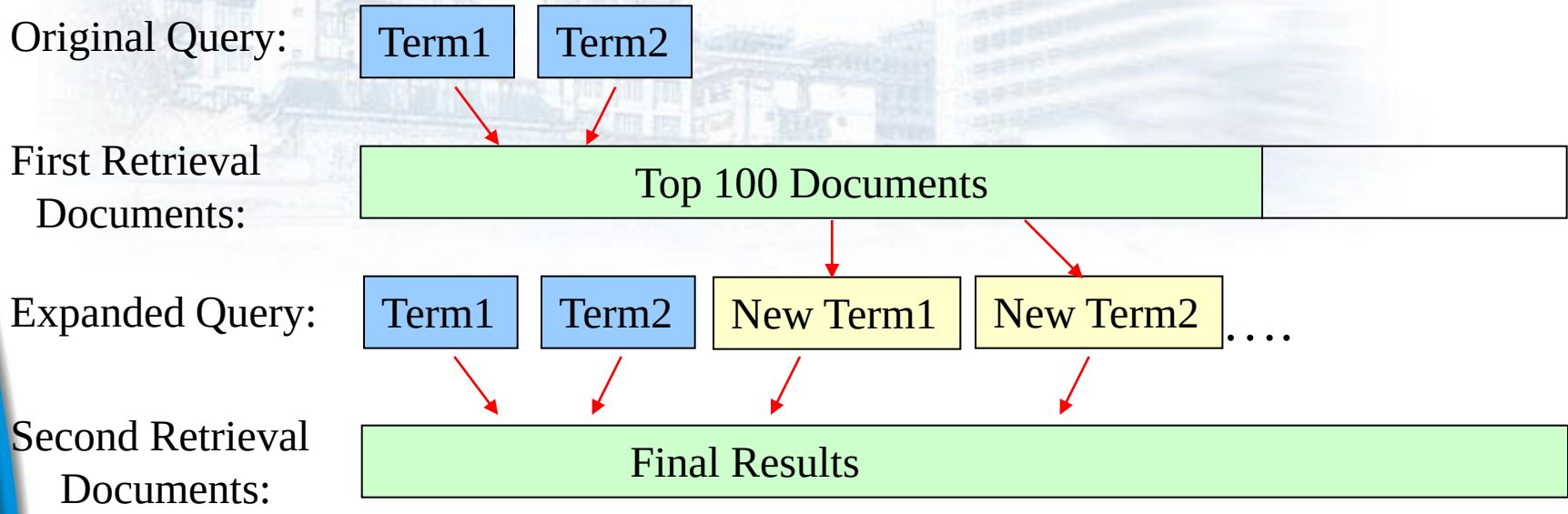
自然地理
（實際管轄區）

首都	加德滿都
最大城市	加德滿都

面積
- 國土面積：140,800平方公里（世界第94名）
- 水域率：2.8%

Pseudo Relevance Feedback

- The pseudo relevance feedback method extracts relevant terms from the result of the first retrieval and uses them as expanded queries to retrieve documents again.



Candidate Weight

- In official run, we used TF-IDF to select 50 expansion terms from Top 100 relevance documents.

$$tf_i = \frac{ni}{\sum_k nk}$$

Ex. term “university” – tf = 1000
df = 800

$$idf = \log \frac{N}{df}$$

term “CYUT” – tf = 100
df = 20

Ranking Method

- The ranking method is Robertson's in our system is the standard OKAPI BM25 algorithm.

$$Sim(Q, D_i) = \sum_{T \in Q} w^1 \frac{(k_1 + 1)tf\,(k_3 + 1)qtf}{(K + tf)(k_3 + qtf)}$$

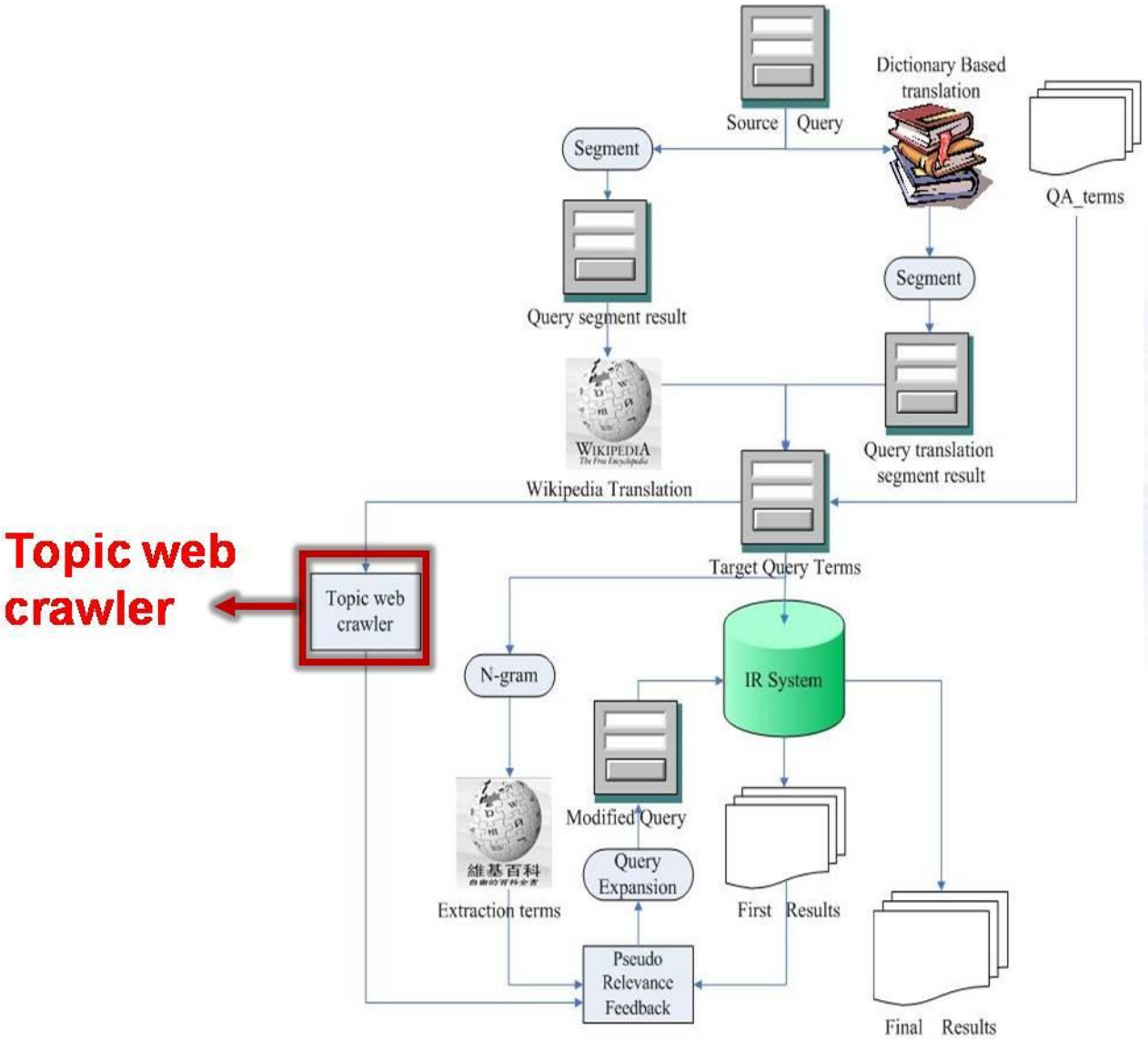
$$w^1 = \log \frac{(r + 0.5)/(R - r + 0.5)}{(n - r + 0.5)/(N - n - R + r + 0.5)}$$

$$K = k_1((1 - b) + b\frac{dl}{avdl})$$

k1=1.2, k3=7, b=0.75

Topic web crawler

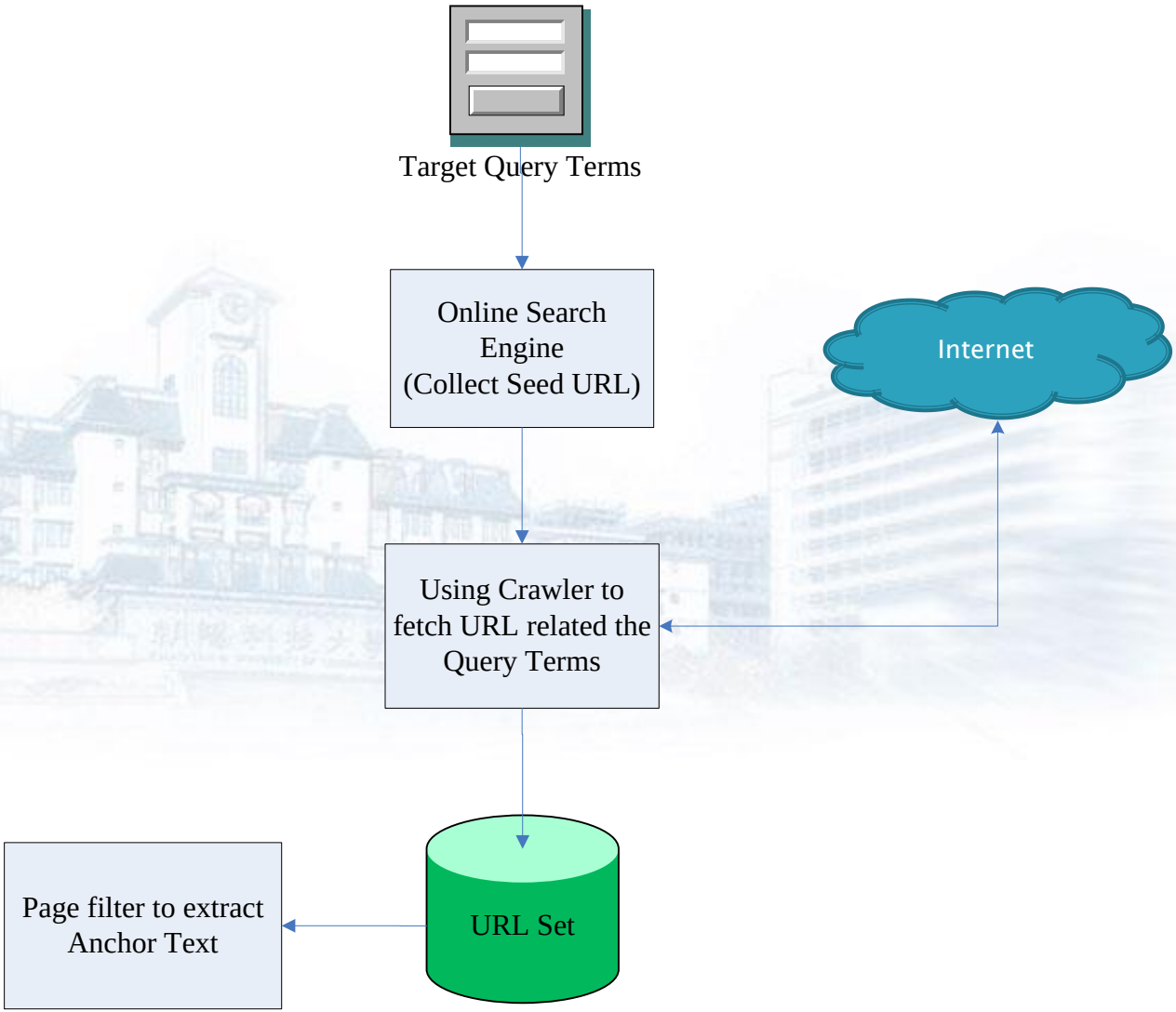
Architecture of retrieval system



Topic web crawler

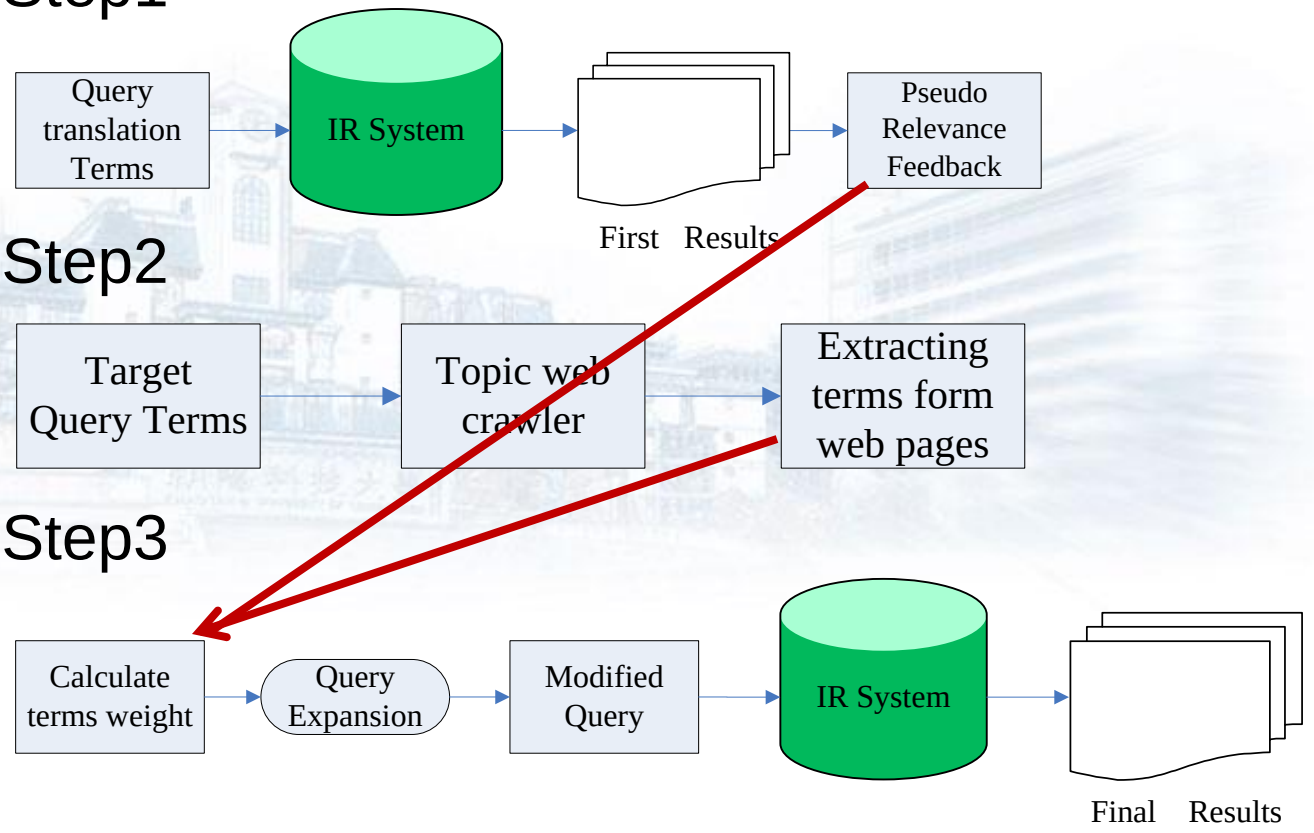
- Topic web crawler is a Web spider program that can retrieve only the documents related to a give topic.
- This kind of crawler is called focused crawler or thematic crawler.
- The key difference of a focused crawler to a general crawler lies on the ability to find more related document among all available links.

Topic web crawler flow chart



Topic web crawler

- Retrieval system flow chart:
  - Step1
  - Step2
  - Step3



Experiment Result

Experiment Result

- We use NTCIR-8 ACLIA IR4QA Subtask data sets run experiment.
  - Document sets:

Language	File name	Number of the docs	Year
Chinese (Simplified)	UDN	308,845	2002-2005
Chinese (Traditional)	Xinhua	1,663,517	2002-2005
Japanese	Mainichi	377,941	2002-2005

Experiment Result

- We use NTCIR-8 ACLIA IR4QA Subtask data sets run experiment.
  - Settings of official runs:

Run Type	Mean
T-run-01	Use only QUESTION field in Topic files as query terms
T-run-02	Adding more terms from answer type analysis of CCLQA to the first setting
D-run	Use the NARRATIVE field in Topic file as the query terms
DN-run	Combine the terms in QUESTION field and NARRATIVE field as the query terms

Official Runs

- The performances of official runs(CS/JA results BEFORE bug fix.)

Run	MAP	M-Q	M-nDCG
EN-CT-T-01	0.1733	0.1923	0.3672
EN-CT-T-02	**0.1941**	**0.2137**	**0.3963**
EN-CT-D-03	0.1362	0.1509	0.321
EN-CT-DN-04	0.1486	0.1667	0.3516
EN-CS-T-01	0.1955	0.2225	0.4152
EN-CS-T-02	**0.1996**	**0.2263**	**0.429**
EN-CS-D-03	0.1445	0.1674	0.3622
EN-CS-DN-04	0.1562	0.1817	0.3933
EN-JA-T-01	0.1708	0.1776	0.3613
EN-JA-T-02	**0.1719**	**0.1788**	**0.3638**
EN-JA-D-03	0.1023	0.1027	0.2565
EN-JA-DN-04	0.0999	0.0985	0.2449

Additional Runs

- In the Additional Runs, we have to process two experiments.
  - Experiment 1: using different proportion in QE term from Okapi and Wikipedia.
  - Experiment 2: using different proportion in QE term Okapi and topic web crawler.
- Through to analysis different proportion in QE term from Okapi and other source, to known which one of QE terms as candidate for query expansion can help CLIR to improve precision.

Additional Runs

- Experiment 1_CS-runs :
  - EN-CS using Okapi QE have better performance.
  - Many T-runs are better than D-runs and DN-runs.
  - Wikipedia QE is helpful in T-Runs.

	Okapi QE : Wikipedia QE(QE term=50)										
Run	**100:0**	**90:10**	**80:20**	**70:30**	**60:40**	**50:50**	**40:60**	**30:70**	**20:80**	**10:90**	**0:100**
CYUT-EN-CS-T	0.2006	0.1984	0.1999	**0.2014**	0.2003	0.1965	0.1948	0.1926	0.186	0.1865	0.1707
CYUT-EN-CS-T(QA)	0.202	0.2014	**0.2031**	0.2028	0.2005	0.2001	0.196	0.1941	0.1894	0.1943	0.1806
CYUT-EN-CS-D	**0.1601**	0.1575	0.1566	0.156	0.1538	0.1472	0.1434	0.1421	0.1386	0.1291	0.1136
CYUT-EN-CS-DN	**0.1696**	0.1668	0.1673	0.1655	0.165	0.1572	0.1565	0.1563	0.1546	0.1489	0.1311

Additional Runs

- Experiment 1_JA-runs:
  - EN-JA using Okapi QE have better performance.
  - Our MAP of the EN-JA run was much lower than the other runs.

	Okapi QE : Wikipedia QE(QE term=50)										
Run	**100:0**	**90:10**	**80:20**	**70:30**	**60:40**	**50:50**	**40:60**	**30:70**	**20:80**	**10:90**	**0:100**
CYUT-EN-JA-T	0.1628	**0.1636**	0.161	0.1603	0.1594	0.1561	0.154	0.1515	0.1428	0.1321	0.1034
CYUT-EN-JA-T(QA)	0.1617	**0.1625**	0.1601	0.1594	0.1583	0.155	0.1528	0.1503	0.1414	0.131	0.1024
CYUT-EN-JA-D	0.0881	0.0928	**0.0929**	0.0917	0.0907	0.0893	0.0877	0.0849	0.0822	0.079	0.058
CYUT-EN-JA-DN	0.0857	0.0904	0.0895	0.0904	**0.0905**	0.0875	0.0851	0.0822	0.0813	0.077	0.0569

Additional Runs

- Experiment 1_CT-runs:
  - EN-CT using Okapi QE have better performance.
  - The MAP of QE terms from Okapi only and Wikipedia only are quite close.
  - Wikipedia QE is helpful in EN-CT.

	Okapi QE : Wikipedia QE(QE term=20)										
Run	**100:0**	**90:10**	**80:20**	**70:30**	**60:40**	**50:50**	**40:60**	**30:70**	**20:80**	**10:90**	**0:100**
CYUT-EN-CT-T	0.1738	0.1738	0.1746	0.1762	**0.1782**	0.1768	0.1752	0.1704	0.1667	0.1648	0.153
CYUT-EN-CT-T(QA)	0.1938	0.1935	0.1943	**0.1948**	0.1971	0.1959	0.1938	0.1911	0.1877	0.1842	0.1697
CYUT-EN-CT-D	0.1382	0.1406	**0.141**	0.1379	0.1395	0.1396	0.1381	0.1352	0.1313	0.123	0.1137
CYUT-EN-CT-DN	0.1559	0.1567	**0.1571**	0.1565	0.1567	0.1555	0.153	0.152	0.149	0.1427	0.1343

Additional Runs

- Experiment 2_CS-runs :
  - EN-CS runs using Okapi QE have better performance.
  - Topic web crawler QE is more helpful in EN-CS.

	Okapi QE : Topic web crawler QE(QE term=20)										
Run	**100:0**	**90:10**	**80:20**	**70:30**	**60:40**	**50:50**	**40:60**	**30:70**	**20:80**	**10:90**	**0:100**
CYUT-EN-CS-T	0.2006	0.205	0.2077	**0.2071**	0.2041	0.1945	0.1949	0.1929	0.1865	0.1846	0.1729
CYUT-EN-CS-T(QA)	0.202	0.2073	0.208	**0.2084**	**0.2084**	0.1998	0.2001	0.1965	0.1932	0.1937	0.1767
CYUT-EN-CS-D	0.1601	0.1638	0.1641	**0.1652**	0.1612	0.1556	0.156	0.1537	0.1496	0.1447	0.1343
CYUT-EN-CS-DN	0.1696	**0.1707**	0.1688	0.1704	0.1681	0.1606	0.1613	0.1623	0.1609	0.159	0.1472

Additional Runs

- Experiment 2_CT-runs
  - Topic web crawler QE is more helpful in T-runs.
  - The MAP of QE terms from Okapi only and Topic web crawler only are quite close in D-runs and DN-runs.

	Okapi QE : Topic web crawler QE(QE term=30)										
Run	**100:0**	**90:10**	**80:20**	**70:30**	**60:40**	**50:50**	**40:60**	**30:70**	**20:80**	**10:90**	**0:100**
CYUT-EN-CT-T	0.1735	0.1769	0.1791	0.1801	0.1816	0.1824	**0.1839**	0.1808	0.1793	0.181	0.1682
CYUT-EN-CT-T(QA)	0.1946	0.1995	0.1999	0.2021	0.2024	0.2044	**0.206**	0.2003	0.1972	0.1974	0.1798
CYUT-EN-CT-D	0.1375	0.1388	0.1431	0.141	0.1457	**0.1461**	0.1449	0.1462	0.1413	0.1409	0.1275
CYUT-EN-CT-DN	0.1566	0.1589	0.1588	0.1614	0.1654	0.1669	**0.1676**	0.1667	0.1651	0.1651	0.1508

Conclusions

Conclusions

- In this paper, we using Wikipedia and Google translation to translate query terms, and using the results of QA analysis to add more target query terms.
- Main of the query expansion terms to extract terms as anchor text from Wikipedia and using topic web crawler extract more keywords to be the candidates of QE.

Conclusions

- In additional runs, the experiment 1 result to show using the Okapi terms can improve the performance of MAP, especially for EN-CS and EN-CT.

- In the experiment 2 result to show using the topic crawler terms better than Okapi terms, therefore to prove topic crawler can help the retrieval system to raise the performance.

Conclusions

- In the future work, because of the question types of the IR4QA task increased from 4 in NTCIR-7 to 9 in NTCIR-8. This change makes the task more difficult.
- Therefore, we think of the IR system must use more information on the question types, such as building classifiers to relate documents to particular question types.

Query Expansion from Wikipedia and Topic Web Crawler on CLIR

Meng-Chun Lin, Ming-Xiang Li, Chih-Chuan Hsu, Shih-Hung Wu

Thank you for your attention!

Proceedings of NTCIR-8 Workshop Meeting, June, 2010

Adviser : Prof. Shih-Hung Wu

Reporter : Meng-Chun Lin