Experiments on Domain Adaptation for Patent Machine Translation in the PLuTO project

Alexandru Ceauşu, John Tinsley,
Jian Zhang, Andy Way

CNGL, Dublin City University

EAMT 2011, May 30, Leuven Belgium

Funded under the EU ICT Policy Support Programme

What is PLuTO?

Patent Language Translations Online

- Online translation (and retrieval) of patent documents
- Machine translation in PLuTO
  - MaTrEx system (hybrid SMT-EBMT)
  - Translation Memory
- TM—MT evaluation
- Languages: English, French, Portuguese
  German, Spanish, Russian, Japanese, Chinese, …
- Patent users translation needs
  - From informal to flawless translation

Funded under the EU ICT Policy SupportProgramme

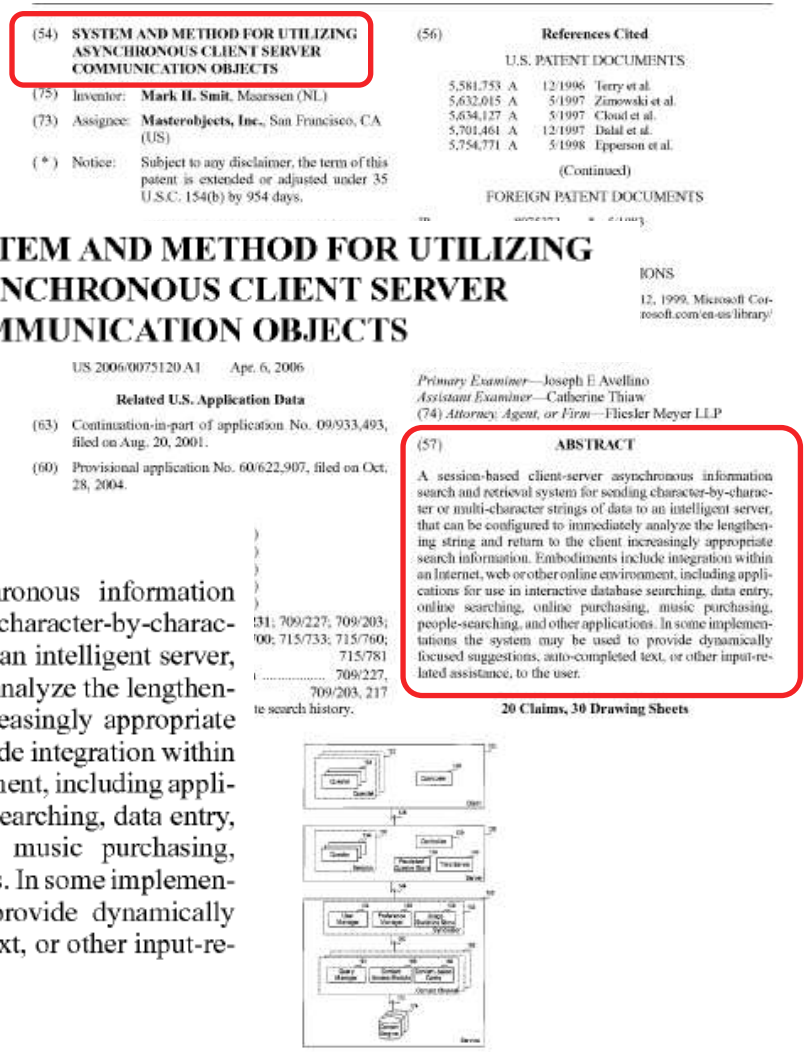
Overview

- Data preparation (English—French)
- Patent domain adaptation for Machine Translation (MT)
  - IPC (International Patent Classification) domains
  - Patent specific formulation and terminology
- Results/findings
- Description of the online MT system

Funded under the EU ICT Policy SupportProgramme

Patent language

- Different formulation requirements for each section of a patent document:
  - *Title*
  - *Abstract*



(54) SYSTEM AND METHOD FOR UTILIZING ASYNCHRONOUS CLIENT SERVER COMMUNICATION OBJECTS

(75) Inventor: Mark H. Smit, Maarssen (NL)

(73) Assignee: Masterobjects, Inc., San Francisco, CA (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 954 days.

(56) References Cited

U.S. PATENT DOCUMENTS

5,581,753 A 12/1996 Terry et al.
5,632,015 A 5/1997 Zimowski et al.
5,634,127 A 5/1997 Cloud et al.
5,701,461 A 12/1997 Dalal et al.
5,754,771 A 5/1998 Epperson et al.

(Continued)

FOREIGN PATENT DOCUMENTS

(54) SYSTEM AND METHOD FOR UTILIZING ASYNCHRONOUS CLIENT SERVER COMMUNICATION OBJECTS

IONS

12, 1999, Microsoft Cor-rosoft.com/en-us/library/

US 2006/0075120 A1 Apr. 6, 2006

Related U.S. Application Data

(63) Continuation-in-part of application No. 09/933,493, filed on Aug. 20, 2001.

(60) Provisional application No. 60/622,907, filed on Oct. 28, 2004.

Primary Examiner—Joseph E Avellino
Assistant Examiner—Catherine Thiaw
(74) *Attorney, Agent, or Firm*—Fliesler Meyer LLP

(57) ABSTRACT

A session-based client-server asynchronous information search and retrieval system for sending character-by-character or multi-character strings of data to an intelligent server, that can be configured to immediately analyze the lengthening string and return to the client increasingly appropriate search information. Embodiments include integration within an Internet, web or other online environment, including applications for use in interactive database searching, data entry, online searching, online purchasing, music purchasing, people-searching, and other applications. In some implementations the system may be used to provide dynamically focused suggestions, auto-completed text, or other input-related assistance, to the user.

20 Claims, 30 Drawing Sheets

(57) ABSTRACT

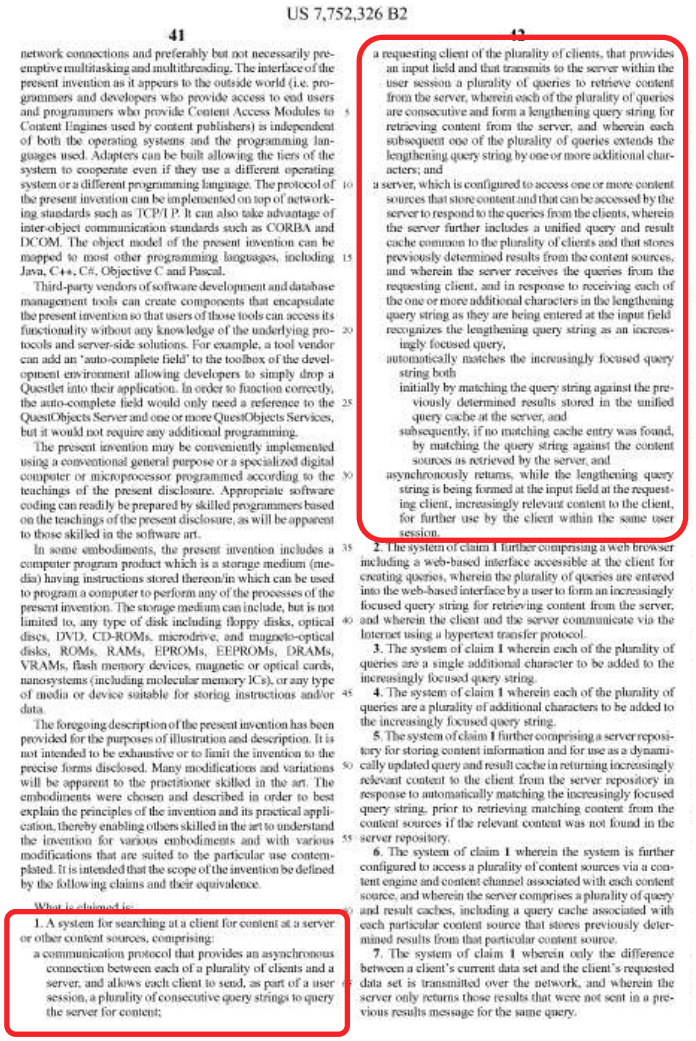
A session-based client-server asynchronous information search and retrieval system for sending character-by-character or multi-character strings of data to an intelligent server, that can be configured to immediately analyze the lengthening string and return to the client increasingly appropriate search information. Embodiments include integration within an Internet, web or other online environment, including applications for use in interactive database searching, data entry, online searching, online purchasing, music purchasing, people-searching, and other applications. In some implementations the system may be used to provide dynamically focused suggestions, auto-completed text, or other input-related assistance, to the user.

31; 709/227; 709/203; 00; 715/733; 715/760; 715/781
..... 709/227, 709/203, 217
te search history.

Funded under the EU ICT Policy SupportProgramme

Patent language

- Different formulation requirements of each section of a patent document:
  - *Title*
  - *Abstract*
  - *Description*
  - *Claims*



US 7,752,326 B2

41

network connections and preferably but not necessarily pre-emptive multitasking and multithreading. The interface of the present invention as it appears to the outside world (i.e. programmers and developers who provide access to end users and programmers who provide Content Access Modules to Content Engines used by content publishers) is independent of both the operating systems and the programming languages used. Adapters can be built allowing the tiers of the system to cooperate even if they use a different operating system or a different programming language. The protocol of the present invention can be implemented on top of networking standards such as TCP/IP. It can also take advantage of inter-object communication standards such as CORBA and DCOM. The object model of the present invention can be mapped to most other programming languages, including Java, C++, C#, Objective C and Pascal.

Third-party vendors of software development and database management tools can create components that encapsulate the present invention so that users of those tools can access its functionality without any knowledge of the underlying protocols and server-side solutions. For example, a tool vendor can add an "auto-complete field" to the toolbox of the development environment allowing developers to simply drop a QuestJet into their application. In order to function correctly, the auto-complete field would only need a reference to the QuestObjects Server and one or more QuestObjects Services, but it would not require any additional programming.

The present invention may be conveniently implemented using a conventional general purpose or a specialized digital computer or microprocessor programmed according to the teachings of the present disclosure. Appropriate software coding can readily be prepared by skilled programmers based on the teachings of the present disclosure, as will be apparent to those skilled in the software art.

In some embodiments, the present invention includes a computer program product which is a storage medium (media) having instructions stored thereon/in which can be used to program a computer to perform any of the processes of the present invention. The storage medium can include, but is not limited to, any type of disk including floppy disks, optical discs, DVD, CD-ROMs, microdrive, and magneto-optical disks, ROMs, RAMs, EPROMs, EEPROMs, DRAMs, VRAMs, flash memory devices, magnetic or optical cards, nanosystems (including molecular memory ICs), or any type of media or device suitable for storing instructions and/or data.

The foregoing description of the present invention has been provided for the purposes of illustration and description. It is not intended to be exhaustive or to limit the invention to the precise forms disclosed. Many modifications and variations will be apparent to the practitioner skilled in the art. The embodiments were chosen and described in order to best explain the principles of the invention and its practical application, thereby enabling others skilled in the art to understand the invention for various embodiments and with various modifications that are suited to the particular use contemplated. It is intended that the scope of the invention be defined by the following claims and their equivalence.

What is claimed is:

1. A system for searching at a client for content at a server or other content sources, comprising:

a communication protocol that provides an asynchronous connection between each of a plurality of clients and a server, and allows each client to send, as part of a user session, a plurality of consecutive query strings to query the server for content;

42

a requesting client of the plurality of clients, that provides an input field and that transmits to the server within the user session a plurality of queries to retrieve content from the server, wherein each of the plurality of queries are consecutive and form a lengthening query string for retrieving content from the server, and wherein each subsequent one of the plurality of queries extends the lengthening query string by one or more additional characters; and

a server, which is configured to access one or more content sources that store content and that can be accessed by the server to respond to the queries from the clients, wherein the server further includes a unified query and result cache common to the plurality of clients and that stores previously determined results from the content sources, and wherein the server receives the queries from the requesting client, and in response to receiving each of the one or more additional characters in the lengthening query string as they are being entered at the input field recognizes the lengthening query string as an increasingly focused query,

automatically matches the increasingly focused query string both

initially by matching the query string against the previously determined results stored in the unified query cache at the server, and

subsequently, if no matching cache entry was found, by matching the query string against the content sources as retrieved by the server, and

asynchronously returns, while the lengthening query string is being formed at the input field at the requesting client, increasingly relevant content to the client, for further use by the client within the same user session.

2. The system of claim 1 further comprising a web browser including a web-based interface accessible at the client for creating queries, wherein the plurality of queries are entered into the web-based interface by a user to form an increasingly focused query string for retrieving content from the server, and wherein the client and the server communicate via the Internet using a hypertext transfer protocol.

3. The system of claim 1 wherein each of the plurality of queries are a single additional character to be added to the increasingly focused query string.

4. The system of claim 1 wherein each of the plurality of queries are a plurality of additional characters to be added to the increasingly focused query string.

5. The system of claim 1 further comprising a server repository for storing content information and for use as a dynamically updated query and result cache in returning increasingly relevant content to the client from the server repository in response to automatically matching the increasingly focused query string, prior to retrieving matching content from the content sources if the relevant content was not found in the server repository.

6. The system of claim 1 wherein the system is further configured to access a plurality of content sources via a content engine and content channel associated with each content source, and wherein the server comprises a plurality of query and result caches, including a query cache associated with each particular content source that stores previously determined results from that particular content source.

7. The system of claim 1 wherein only the difference between a client's current data set and the client's requested data set is transmitted over the network, and wherein the server only returns those results that were not sent in a previous results message for the same query.

Funded under the EU ICT Policy SupportProgramme

Patent language

- MT challenges in translating patent discourse:
  - inconsistent terminology

    (54) **SYSTEM AND METHOD FOR UTILIZING ASYNCHRONOUS CLIENT SERVER COMMUNICATION OBJECTS**
  - long sentences

    Claims are expressed in a single sentence.
  - references to figures

    Preferably , there is more than one leg (16 , 17 , 18) that is attached to the bottom of the base member (12) .
  - chemical compounds

    L is an organic group selected from -CH2-(OCH2CH2)n-, -CO-NR'-, with R'=H or C1-C4 alkyl group; n=0-8; Y=F, CF3 …
  - numerical expressions

    maximum stress of 1.2 to 3.5 N/mm<2> and a maximum elongation of 700 to 1,300% at 0[deg.] C.

Funded under the EU ICT Policy SupportProgramme

Data preparation

Data preparation is a *crucial* task when it comes to patent MT:

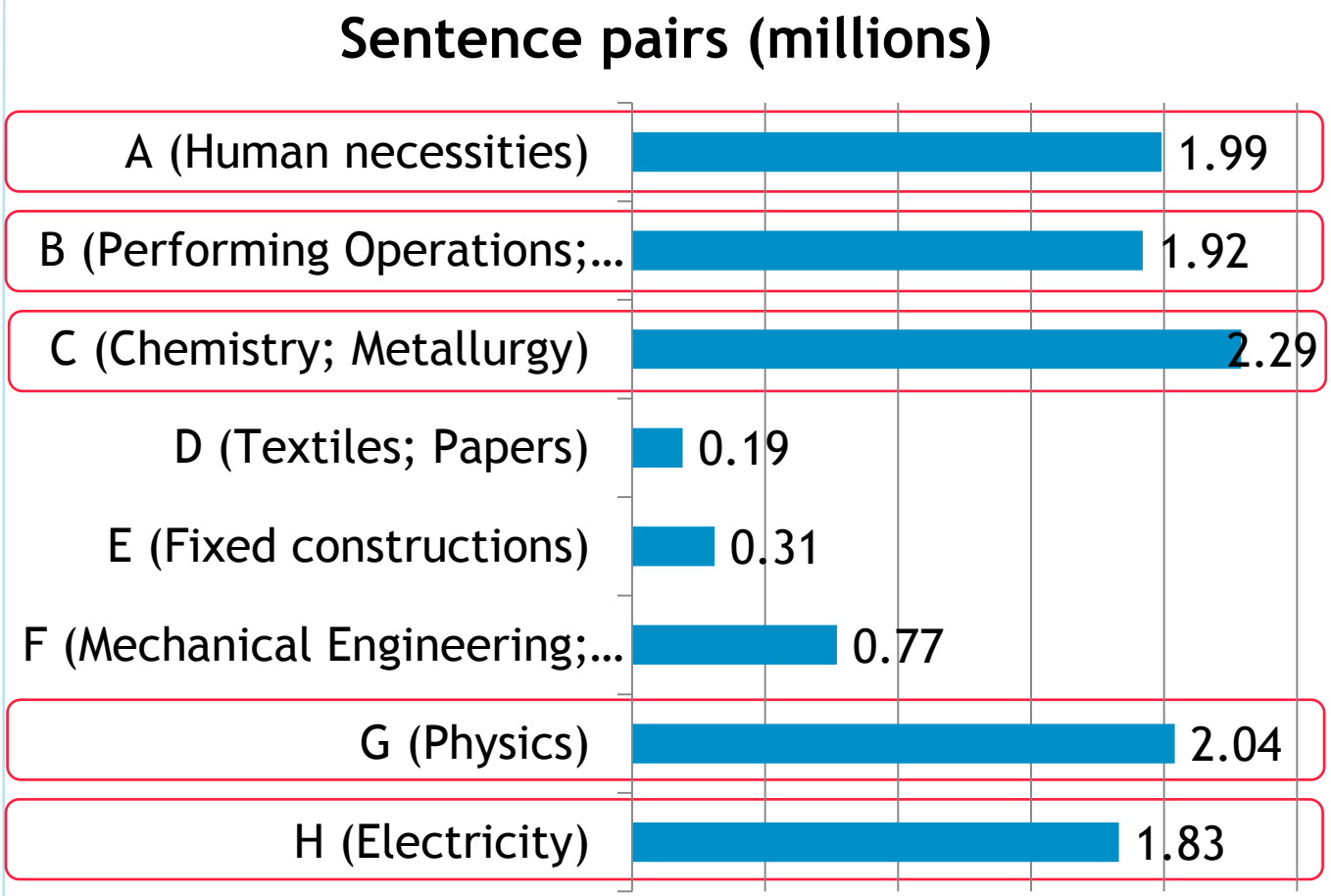
- Pre-processing (training and runtime)
  - Casing and tokenisation (special parsing for compounds, formulae)
  - Handling of lists, references, and figures
  - Splitting long sentences
- Post-processing
  - Reintroduction of case, references, etc.
  - Reassembly of output sentence

MAREC-IRF corpus (more than 650 GB of multi-lingual patent documents)

	English	French	English–French
Abstract	16.57 M	1.68 M	1.65 M
Claims	14.91 M	7.70 M	7.56 M
Description	7.85 M	0.20 M	0 M

Funded under the EU ICT Policy SupportProgramme

Training data



		English tokens	French tokens
A (Human necessities)	1.99	65 M	74 M
B (Performing Operations;…	1.92	71 M	79 M
C (Chemistry; Metallurgy)	2.29	70 M	79 M
D (Textiles; Papers)	0.19	6 M	7 M
E (Fixed constructions)	0.31	11 M	13 M
F (Mechanical Engineering;…	0.77	29 M	33 M
G (Physics)	2.04	68 M	78 M
H (Electricity)	1.83	63 M	72 M
Total	**11.39 M**	**387 M**	**438 M**

Funded under the EU ICT Policy SupportProgramme

IPC domain adaptation

- We can adapt in different areas of the MT system:
  - Language models (in-domain or general)
  - Translation models (in-domain or general)
  - Domain specific dictionaries

- Previous findings
  - Some domains are more compatible than others
  - Domain adaptation is contingent on the amount of available data

Tinsley et al, *PLuTO: MT for Online Patent Translation*, AMTA 2010

Funded under the EU ICT Policy SupportProgramme

Domain adaptation evaluation

English—French	In-domain TM, In-domain LM	In-domain TM, General LM	General TM, In-domain LM	General LM, General TM
All				56.28 / 65.45
A (Human necessities)	56.81 / 65.52	57.18 / 65.81	55.59 / 64.41	56.21 / 65.45
B (Operations)	55.75 / 65.54	56.31 / 65.90	54.59 / 64.45	55.57 / 65.76
C (Chemistry)	59.73 / 68.52	59.93 / 68.58	58.96 / 67.98	60.90 / 69.18
G (Physics)	54.97 / 65.61	55.18 / 65.73	54.58 / 64.90	54.74 / 65.32
H (Electricity)	55.30 / 65.50	55.76 / 65.83	54.47 / 64.85	55.18 / 65.61

BLEU / METEOR score

Funded under the EU ICT Policy SupportProgramme

Patent language adaptation

- Reordering and translation constraints
  - Chemical compounds
  - Numerical expressions
  - References to figures
  - Bullet points
- Input segmentation
  - MaTrEx marker based chunker for input segmentation

Preferably , there is more than one leg (14 , 17 , 18) that is attached to the bottom of the base member (12) .

L is an organic group selected from <chem translation="-CH2-(OCH2CH2)n-, -CO-NR'-">-CH2-(OCH2CH2)n-, -CO-NR'-</chem>, with R'=H or <chem translation="groupe alkyle en C1-C4">C1-C4 alkyl group</chem>; n=0-8; Y=F, CF3 …

A device according to any preceding claim ,

further com-prising illumination means for illuminating the eye of said user ,

wherein said viewpoint detecting means is adapted to detect said viewpoint

by receiving the light emitted by said illumination means and reflected by the surface of said eye .

Funded under the EU ICT Policy SupportProgramme

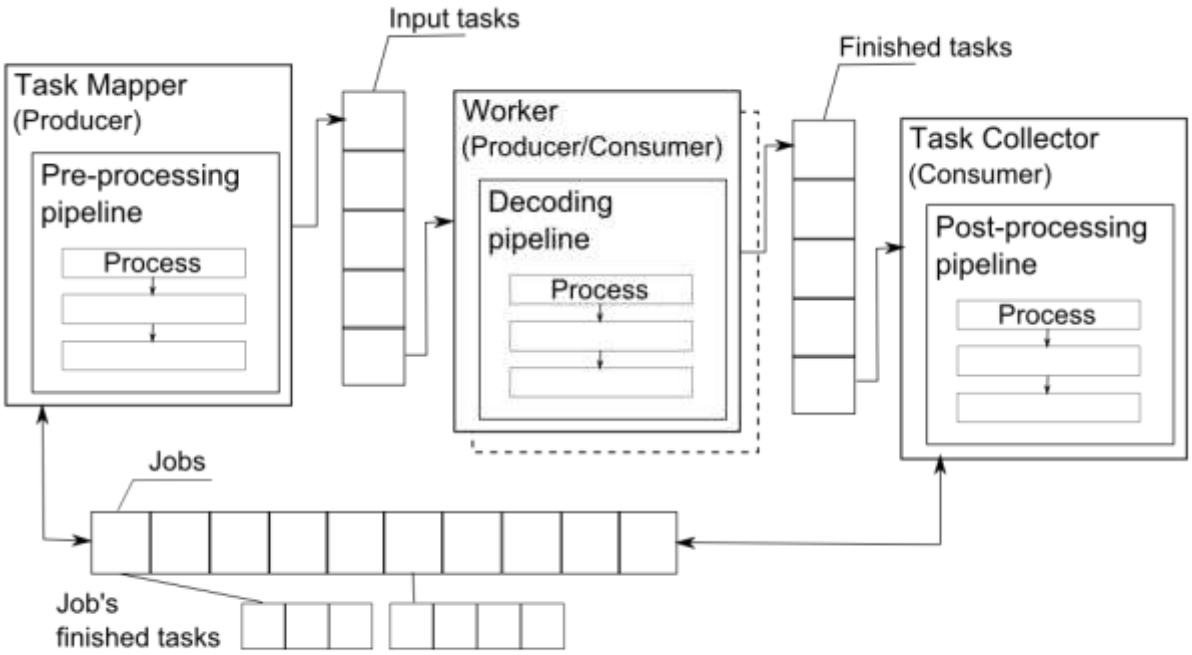
Automatic evaluation

English-French MT System	BLEU	METEOR-NEXT
PLuTO	56.95	66.32
Google	42.67	57.00
Systran	31.62	50.12

French-English MT System	BLEU	METEOR-NEXT
PLuTO	56.92	67.90
Google	42.52	59.65
Systran	28.90	53.67

Funded under the EU ICT Policy SupportProgramme

MT Web Service



PLuTO MT server – multiple producers-consumers architecture

- MT was exposed as a web service to facilitate integration with EspaceNet, the prototype system, and other features
- Job scheduling architecture allows of efficient processing of multiple translations in parallel

Funded under the EU ICT
Policy SupportProgramme

Enter Sentence

A potassium sulfate producing process includes the reaction of sodium sulfate or mirabilite, potassium chloride and system mother liquor in the weight ratio of (1.7-2.15) to (0.08-0.20) to (8.19-10.76) and solid-liquid separation to obtain glaserite and salt making mother liquor, the reaction of glaserite, potassium chloride and water in the weight ratio of (1.1-1.2) to (0.65-0.95) to (1.6-1.9) on potassium sulfate and potassium other liquor. The present invention has the advantages of short technological process, short reaction period, low power consumption, high product yield and quality, and no "three waste" discharge.

Source Language: English

Target Language: French

Translate

A potassium sulfate producing process includes the reaction of sodium sulfate or mirabilite, potassium chloride and system mother liquor in the weight ratio of (1.7-2.15) to (0.08-0.20) to (8.19-10.76) and solid-liquid separation to obtain glaserite and salt making mother liquor, the reaction of glaserite, potassium chloride and water in the weight ratio of (1.1-1.2) to (0.65-0.95) to (1.6-1.9) on potassium sulfate and potassium other liquor. The present invention has the advantages of short technological process, short reaction period, low power consumption, high product yield and quality, and no "three waste" discharge.	Un procédé de production de sulfate de potassium comprend la réaction de mirabilite ou le sulfate de sodium, chlorure de potassium et système de liqueur mère dans le rapport pondéral de (1.7-2.15) à (0.08-0.20) à (8.19-10.76) et de séparation solide-liquide pour obtenir glasérite et sel de fabrication de liqueur mère, la réaction de glasérite, du chlorure de potassium et de l'eau dans le rapport pondéral de (1.1-1.2) à (0.65-0.95) à (1.6-1.9) sur du sulfate de potassium et d'autres liqueur. La présente invention présente les avantages suivants: court processus technologique, de réaction courte période, une faible consommation d'énergie, à rendement élevé et de qualité, et aucun "trois" d'évacuation des

Funded under the EU ICT Policy SupportProgramme

Future plans for MT in Pluto

- MT engines for more languages:
  - Focus on the demand: German, Russian, Asian languages
- Improve integration with the Translation Memory component
- More efficient language and translation models
- Development of user interfaces based on feedback
  - From our partner WON and other users we have engaged with

Funded under the EU ICT Policy SupportProgramme

Thank you

Poster

- 10 AM tomorrow

Funded under the EU ICT Policy SupportProgramme