Employment & Unemployment Insights from Search Data

Baidu Index + Baidu Baike: Acquisition, Index Construction, Descriptive Analysis

Chenxi Zhang Haowen Shi Haotian Zhou

Data Science & AI Course Project (Project B)

December 25, 2025

Outline

Background

- Employment/unemployment is closely tied to social stability and macro conditions.
- Search interest provides a high-frequency **proxy** for public concern and job-market sentiment.
- We combine **Baidu Index** (behavioral signal) with **Baidu Baike** (concept/policy context) for interpretability.

Project Objective

Construct an interpretable **labor-market attention index** and summarize descriptive findings (trend, distribution, robustness).

Research Questions

1. How does search interest in employment/unemployment topics evolve over time?
2. Can multiple keywords be aggregated into meaningful **sub-indices** and a **composite index**?
3. Are the main patterns stable under different normalization and aggregation choices?

Midterm Plan: Scope & Deliverables

Scope (planned)

China, past five years (2019–2025).

Deliverables (planned)

Clean datasets, descriptive analytics, and clear visuals; reproducible code + report/slides.

Pipeline (planned)

- Data Collection (Baidu Index & Baidu Baike)
- Data Cleaning (missingness/duplicates/outliers)
- Descriptive Analysis (distribution/correlation/trends)
- Visualization (time trends, comparisons)
- Conclusions & Outlook

Midterm Plan: Timeline & Team Roles

Timeline (next 4–5 weeks)

- Week 1: finalize keyword list; collect Index & Baike; build raw database
- Week 2: cleaning pipeline; missing/outlier handling; unify region codes
- Week 3: descriptive analysis; correlation/trends; event annotations
- Week 4: visualization polishing; interpret results; draft report/slides
- Week 5 (optional): forecasting / policy comparison

Roles & Allocation

- Chenxi Zhang: keyword design; Index export; regional comparison
- Haowen Shi: Baike text structuring; text mining/topic sketches
- Haotian Zhou: cleaning scripts; time-series plots; robustness checks

Data Sources

Baidu Index (Primary quantitative source)

Keyword-level search index time series (configurable by time window and region).

Baidu Baike (Contextual source)

Definitions / categories / policy background used to justify keyword selection and grouping.

Current dataset used in this report

Monthly series from 2023-01 to 2024-12 (24 timestamps), 42 keywords across 4 sub-domains.
Note: overlapping keywords (e.g., 临时工、兼职、蓝领招聘) appear in multiple sub-domains by design.

Data Acquisition

Baidu Index collection

- Define keyword list and sub-domain mapping.
- Export index series for a chosen time window (and region if needed).
- Store into structured tables for downstream processing.

Baidu Baike collection

- Scrape targeted entries (policies, unemployment types, etc.).
- Extract structured text fields to support interpretability and later text mining.

Data Cleaning

- **Completeness**: check missing timestamps/keywords; interpolate or drop with records.
- **Duplicates**: ensure uniqueness within each (Timestamp, Keyword, Domain) key.
- **Outliers**: flag abnormal spikes (3σ/IQR); keep if event-driven.
- **Formatting**: unify date formats and numeric types; consistent encoding.

Normalization

Why normalize?

Keyword series have different scales; normalization prevents dominance by large-scale keywords.

Two baseline methods

MinMax: $x' = \dfrac{x - \min(x)}{\max(x) - \min(x)}$ Z-score: $x' = \dfrac{x - \mu}{\sigma}$

Sub-domains & Keyword Design (Updated)

Four sub-domains (updated after instructor feedback)

1. 求职活跃度 (Job-search & Hiring Demand)
2. 就业困难/就业压力 (Employment Difficulty / Uncertainty)
3. 失业/裁员/再就业压力 (Unemployment & Layoff Pressure)
4. 结构性/弱势群体就业压力 (Structural / Precarious Employment Stress)

Notes on overlapping keywords

Some terms (e.g., “兼职/临时工/蓝领招聘”) can reflect both general labor demand and structural stress. We keep the mapping explicit and use robustness checks to test alternative assignments.

Keyword List (Final)

求职活跃度：找工作，招聘，求职，招聘信息，校招，春招，秋招，面试，简历

就业困难/压力：就业难，找工作难，应届生就业，应届生找工作，就业形势，就业前景，行业前景，薪资水平，薪资查询，大学生就业，毕业生就业

失业/裁员压力：失业，裁员，裁员潮，被裁，优化，失业金，失业保险，失业补助，失业登记，再就业，低门槛工作，临时工，兼职，蓝领招聘

结构性/弱势群体压力：35 岁就业，35 岁找工作，中年就业，蓝领招聘，低学历就业，外卖骑手，快递员，送外卖，兼职，临时工，底层劳动岗位

Sub-indices and Composite Index

Sub-index (group average)

For sub-domain d with keyword set $\mathcal{K}_d$:

$$S_d(t) = \frac{1}{|\mathcal{K}_d|} \sum_{k \in \mathcal{K}_d} x'_k(t)$$

Composite index (baseline)

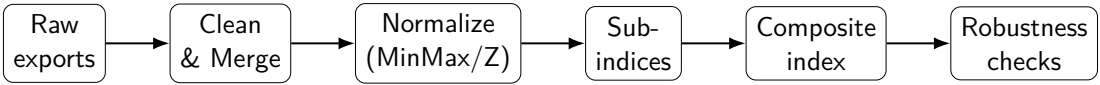
$$I(t) = \sum_{d=1}^{4} w_d \, S_d(t), \qquad \sum_{d=1}^{4} w_d = 1$$

Baseline: equal weights. Robustness: Z-score normalization / MPI aggregation / alternative weights.

Current Sub-index Snapshot (t = 2023-01)

Sub-domain	Interpretation	Value
求职活跃度	job-search intensity	6.88
就业困难/压力	perceived difficulty/uncertainty	18.76
失业/裁员压力	unemployment/layoff concern	70.63
结构性/弱势群体压力	structural/precarious stress	62.87

Processing Pipeline



Reproducible workflow implemented in the repository scripts and exported tables.

Descriptive Summary (Composite Index)

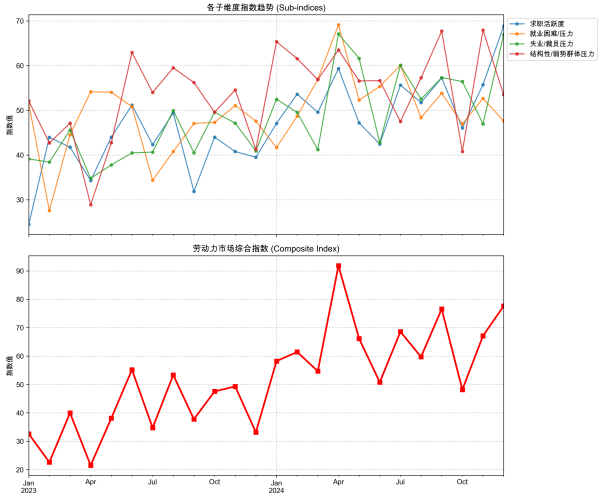
Basic statistics (2023-01 to 2024-12)

- Mean: 36.57 Std: 19.47
- Min: 6.12 Max: 71.53
- Outliers (3σ rule): none detected

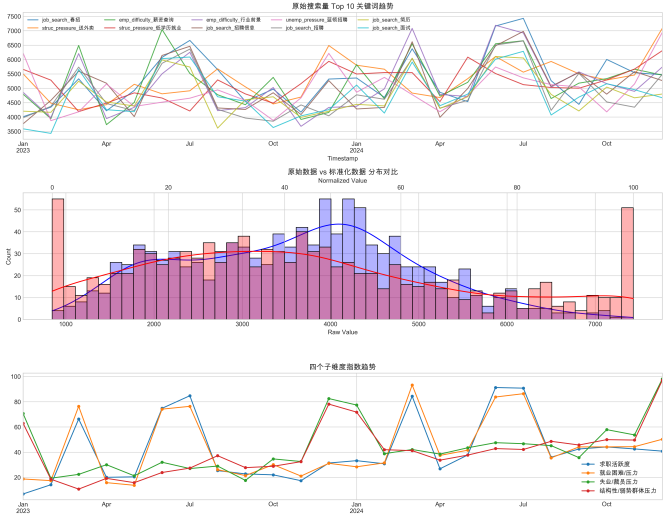
Peak / trough months (examples)

- Top-3 peaks: 2024-12 (71.53); 2024-07 (65.83); 2024-06 (65.61)
- Bottom-3 troughs: 2023-02 (6.12); 2023-05 (6.63); 2023-04 (10.90)

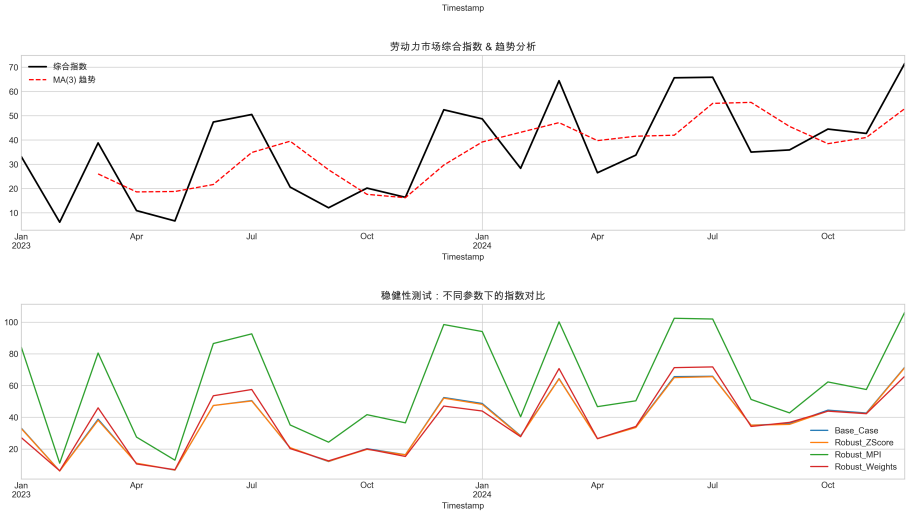
Result Figure: Labor Market Index Analysis



Result Figure: Report Snapshot (Top)



Result Figure: Report Snapshot (Bottom)



Robustness Check (Alternative Constructions)

What we compare

- Base Case vs Z-score normalization
- Linear aggregation vs MPI (non-compensatory aggregation)
- Equal weights vs alternative weights

Correlation (higher means more consistent)

	Base	ZScore	MPI	Weights
Base	1.000000	0.999942	0.939957	0.982355
ZScore	0.999942	1.000000	0.939331	0.982513
MPI	0.939957	0.939331	1.000000	0.917450
Weights	0.982355	0.982513	0.917450	1.000000

Limitations

- Search attention is a **proxy**, not official unemployment/employment measurement.
- News/viral events may create temporary spikes unrelated to structural changes.
- Keyword choice and mapping may introduce subjectivity $\rightarrow$ sensitivity analysis is necessary.

Conclusion & Next Steps

Conclusion

- Built a reproducible pipeline from Baidu Index keyword series to 4 sub-indices and a composite index.
- Produced descriptive summaries and visualizations of labor-market attention dynamics.
- Verified stability under multiple construction choices (normalization/aggregation/weights).

Next Steps (from plan)

Extend to full 2019–2025 window; enrich Baike text mining (TF-IDF/topic sketches); add event and regional comparisons; optional external validation with official statistics.

Q & A